UNIVERSIDAD NACIONAL DE PIURA

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE ESTADÍSTICA

MODELO LOGÍSTICO BINARIO PARA IDENTIFICAR FACTORES DE RIESGO ASOCIADOS A LAS ENFERMEDADES NO TRANSMISIBLES (HIPERTENSIÓN ARTERIAL Y DIABETES MELLITUS) EN LA POBLACIÓN DEL DEPARTAMENTO DE PIURA EN EL AÑO 2013

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE LICENCIADO EN ESTADÍSTICA

AUTOR

Br. KEYLA KARYN MURILLO AREVALO

ASESOR

Dr. CARLOS EDUARDO CABRERA PRIETO

PIURA – PERÚ

2015

UNIVERSIDAD NACIONAL DE PIURA

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE ESTADÍSTICA

Br. KEYLA KARYN MURILLO AREVALO
AUTOR

Dr. CARLOS EDUARDO CABRERA PRIETO

ASESOR

JURADO DE TESIS:

Dr. CONRRADO SIGIFREDO VARGAS LYNCH
PRESIDENTE

MSc. ANA MARILU LEON SILVA
SECRETARIO

Lic. LEMIN ABANTO CERNA
VOCAL

RESUMEN

Las enfermedades no transmisibles en el departamento de Piura han tenido un crecimiento considerable en los últimos años debido a la no importancia y cuidado de los factores de riesgo que la afectan y que se conocen clínicamente por investigaciones en medicina y estudios de salud, por eso una aplicación importante de la ciencia estadística en esta rama de especialización conocida en el contexto académico como bioestadística, es asumir la realización de investigaciones donde se emplee los modelos de regresión para determinar factores de influencia en ciertas enfermedades y técnicas de análisis que ayuden a demostrar, comprobar e identificar características específicas en temas o estudios de salud, comprobables teóricamente y que puedan ser validadas significativamente por los procedimientos estadísticos. El objetivo de esta investigación es determinar un modelo Logístico binario Multivariado que permita calcular la probabilidad de pertenencia o existencia de las enfermedades Diabetes Mellitus e Hipertensión Arterial en la población del departamento de Piura, con ayuda de la información registrada en las bases de datos que le pertenece al INEI, dicha información que es registrada después de ser recolectada por su personal de campo en los estudios de salud, que realiza constantemente en los últimos años. El resultado de la investigación muestra la metodología empleada, los ajustes necesarios y las variables escogidas para el proceso del diseño, finalmente se logra obtener dos modelos con un buen ajuste estadístico aunque con pocos factores de los que inicialmente se ingresaron. Los modelos obtenidos son un gran paso al universo de requerimientos que se deben considerar para desarrollar este tipo de técnicas de regresión, sabiendo que la investigación nuestra es una etapa inicial en este contexto académico y por lo tanto se debe empezar a realizar modelos parecidos para establecer comparaciones, intercambiar conocimientos e identificar procedimientos similares que signifiquen ayuda y material de consulta en el futuro. Teniendo en cuenta la consideración paralela de nuestra investigación donde se hace énfasis en constituir las metodologías de investigación y estadísticas propias de la carrera profesional de la cual formamos parte; desde el enfoque teórico y práctico en consecuencia de que se cuenta con disponibilidad de material en cuanto a los fundamentos y conceptos relevantes de la técnica utilizada y además propia de nuestra especialidad para describir aquellas referencias sobre este tema puntual de nuestro título de investigación, con el cual se pretende marcar la base principal de los modelos logísticos en esta parte de nuestro país, dado que solo se cuenta con investigaciones de

tipo descriptivas bajo estos títulos similares. La descripción de las variables del modelo se muestra en el capítulo de Metodología de la Investigación y por último los resultados obtenidos, su justificación y conclusiones respectivas se muestran en la parte final de este trabajo.

Palabras clave: Hipertensión Arterial, Diabetes Mellitus, Probabilidad, Prevalencia, Factores de Riesgos, Logit, Regresión Logística Binaria, Sensibilidad, Especificidad.

ABSTRACT

Noncommunicable diseases in the department of Piura have had considerable growth in recent years due to the non-importance and care of the risk factors that affect and are clinically known for research in medicine and health studies, so a important application of statistical science in this field of specialization known in the academic context and biostatistics, is to assume the conduct of research where regression models were used to determine factors influencing certain diseases and analysis techniques that help demonstrate check and identify specific features on issues or health studies, theoretically verifiable and can be validated by statistical methods significantly. The objective of this research is to determine a binary multivariate logistic model that calculates the probability of belonging or existence of diseases Diabetes Mellitus and Hypertension in the city of Piura, using the information recorded in databases that belongs to INEI, the information that is recorded after being collected by field staff in health studies, which performs consistently in recent years. The result of the research shows the methodology used, the necessary adjustments and variables chosen for the design process, finally manages to get two good statistical models, but few factors that initially admitted setting. The models obtained are a big step into the world of requirements that must be considered in developing this type of regression techniques, knowing that our research is an initial step in this academic context and therefore must begin making similar models to establish comparisons, exchange knowledge and identify similar procedures that involve help and reference material in the future. Given the parallel consideration of our research where the emphasis is on research methodologies establish and own career statistics which are part; from the theoretical and practical approach accordingly that it has availability of material in terms of the fundamentals and concepts relevant to the technique used and also own our specialty to describe those references on this timely topic of our research degree, in which is intended to make the main base of the logistic models in this part of our country, since it only has descriptive research type under these same title. The description of the model variables shown in the chapter on Research Methodology and finally the results, its rationale and respective conclusions are shown in the final part of this work.

Keywords: Hypertension, Diabetes Mellitus, Probability, Prevalence, Risk Factors, Logit, Binary Logistic Regression, Sensitivity, Specificity.

DEDICATORIA

A Dios, por permitirme llegar a este momento en mi vida. Por los triunfos y pruebas difíciles que me han enseñado a valorar la vida cada día más, y por la oportunidad cada día de tener una esperanza de encontrar el amor, el perdón, y la paz divina que se halla solo en la fe en su hijo, Cristo Jesús, redentor y Salvador. A mi madre, que a pesar de su ausencia es el pilar más importante en mi vida. A mi padre, por demostrarme siempre su cariño e hizo todo en la vida para que yo pudiera lograr mis sueños, por motivarme y darme la mano cuando sentía que el camino se terminaba.

Keyla Karyn Murillo Arévalo.

AGRADECIMIENTO

Le agradezco a Dios por haberme acompañado y guiado a lo largo de mi carrera, por ser mi fortaleza en los momentos de debilidad y por brindarme una vida llena de aprendizajes, experiencias y sobre todo felicidad.

Le doy gracias a mis padres Rodolfo y Cristina por apoyarme en todo momento, por los valores que me han inculcado, y por haberme dado la oportunidad de tener una excelente educación en el transcurso de mi vida. Sobre todo por ser un excelente ejemplo de vida a seguir. A mis hermanas por ser parte importante en mi vida y por representar la unidad familiar.

Agradezco al Dr. Carlos Eduardo Cabrera Prieto, docente de la especialidad, por haber aceptado ser mi asesor, por su gran amistad, apoyo y guiarme en cuanto a la realización de este proyecto.

A Antonio, por ser una parte muy importante de mi vida, por haberme apoyado en las buenas y en las malas, sobre todo por su paciencia y amor incondicional.

Les agradezco la confianza, apoyo y dedicación de tiempo a mis profesores que fueron parte de mi formación profesional.

Gracias al Instituto Nacional de Estadística e Informática (INEI- ODEI PIURA), por haberme brindado la oportunidad de desarrollar mi tesis y la facilidad de contar con la información que recolecta la Encuesta Demográfica y de Salud Familiar – ENDES. Por darme la oportunidad de crecer profesionalmente y aprender cosas nuevas.

Keyla Karyn Murillo Arévalo.

INDICE

I. GENERALIDADES

1.1. INTRODUCCIÓN

Las enfermedades crónicas no transmisibles, afectan a todos los grupos de edad y constituyen un grupo heterogéneo de padecimientos como la diabetes e hipertensión arterial, entre otros; contribuyendo a un problema de salud pública por ser una causa de morbilidad, en el marco del proceso de envejecimiento de la población en nuestro país y por el modo de vida poco saludable (OMS, 2010, pág. i – 1).

Las enfermedades crónicas son aquellas enfermedades de larga duración, degenerativas, no transmisibles mediante el contacto personal y generalmente, interfieren con la capacidad del cuerpo para funcionar de manera óptima. Estas enfermedades no son curables pero pueden ser prevenidas o evitadas con intervenciones para reducir los factores de riesgo que conducen al desarrollo de las mismas por ejemplo, evitando o minimizando el consumo de tabaco, del alcohol, aumentando la actividad física y consumiendo una dieta balanceada; estas consideraciones están establecidas de acuerdo a criterios médicos basados en su experiencia, que por lo tanto desean ser corroborados para poder definir una afirmación o negativa en función de los resultados que se puedan encontrar en esta investigación.

Dentro de las enfermedades crónicas, son de citar: Diabetes, colesterol y/o triglicéridos altos, enfermedades renales, asma, cáncer, enfermedades cardiovasculares, hipertensión arterial, entre otras; siendo las principales enfermedades no transmisibles y detalladas en este estudio, la Hipertensión Arterial y Diabetes Mellitus. Según la Organización mundial de la Salud, las enfermedades crónicas son responsables del 60% de la mortalidad a nivel mundial y su prevalencia es mayor en la población adulta, situación en la que están inmersos los países en vías de desarrollo como el Perú. Con tal motivo, el Ministerio de Salud ha implementado la **Estrategia Sanitaria Nacional de**

Prevención y Control de Daños No Transmisibles, a fin de mediar una acción preventiva que modifique el nivel de los factores de riesgo de estas enfermedades entre la población de nuestro país y disminuir la morbilidad y mortalidad causadas por estas enfermedades.

En los últimos años, los estudios sobre Enfermedades No Transmisibles se han incrementado, en dicho contexto, desde el 2010 la Encuesta Demográfica y de Salud Familiar - ENDES ha incorporado una batería de preguntas con la finalidad de captar y proveer información actualizada sobre la población afectada, para la evaluación y formulación de programas de salud orientados a disminuir o atenuar la prevalencia de las mismas (41% de la población de 60 años a más a nivel nacional tienen presión arterial alta, de acuerdo al ENDES 2013). Pero, la mayoría de estos estudios son descriptivos, ninguno ha analizado la relación estadística de asociación entre las Enfermedades No Transmisibles y las variables que la explican.

La presente investigación consiste en conocer cuáles son los factores de riesgo asociados al comportamiento de las personas que influyen en la existencia de las enfermedades no transmisibles (Hipertensión Arterial y Diabetes Mellitus) en la población del departamento de Piura, que se pueden predecir con la utilización de un modelo Logístico Binario. Dada una significancia estadísticamente valida, y poder asumir aquellos factores como características de riesgo, en la prevalencia de estas enfermedades, teniendo en cuenta que el modelo logístico desarrollado, busca comprobar la influencia o no de dicho factores, dado que por condiciones médicas o evaluaciones de salud relacionadas a las enfermedades, se sabe que los factores incluidos casi siempre condicionan la prevalencia de las enfermedades estudiadas en las personas de nuestra población objetivo.

En el ámbito propio de la investigación estadística este trabajo toma en cuenta el orden que se debe seguir en la utilización de los modelos de regresión logística que están descritos en el contenido de esta investigación, con el objetivo de proporcionar una fuente de consulta en investigaciones parecidas o referentes en

el mismo campo de estudio que se lleguen a realizar o explorar. Otra de las consideraciones de referencia en cuanto al contenido de la investigación es que aquí se tiene las diferencias teóricas de acuerdo a cuando y como se debe realizar regresión logística binaria, los fundamentos que explican este modelo en este trabajo de investigación son importantes para aplicar conceptos básicos de estadística como exploración de datos, tipos de variables, aleatoriedad y normalidad, entre otros conceptos fundamentales. Para llevar a la practica con una sólida base teórica de los conocimientos impartidos a lo largo de nuestra carrera profesional. Finalmente en los resultados de esta investigación se muestra el modelo obtenido y se realizan las comparaciones aquí mencionadas.

La investigación se ha basado en el análisis de la Encuesta Demográfica y de Salud familiar (ENDES) que efectúa el INEI, puesto que esta encuesta tienen información a nivel nacional y ofrecen una muestra muy amplia de casos en los cuales se puede obtener todos los beneficios del análisis estadístico. Las ENDES usada corresponden al departamento de Piura del año 2013. La aplicación inmediata de esta investigación es establecer un grupo de resultados propios del análisis de regresión logística binaria que sean propicios para considerarlos como base o punto de partida de los factores de salud implicados en relación con las enfermedades incluidas en esta investigación.

1.2. EL PROBLEMA DE INVESTIGACIÓN

1.2.1. Descripción del Problema

Las enfermedades crónicas, también conocidas como enfermedades no transmisibles (ENT), principalmente las enfermedades cardiovasculares, el cáncer, enfermedades respiratorias crónicas y la diabetes, son la mayor causa de muerte prematura y de discapacidad en la mayoría de los países de las Américas. Estas enfermedades comparten factores de riesgo comunes que incluyen el tabaquismo, la inactividad física, el uso nocivo del alcohol y la dieta no saludable. Las ENT se pueden prevenir y controlar a través de

cambios en el estilo de vida, políticas públicas e intervenciones de salud, y requieren un abordaje intersectorial e integrado.

Se estima que casi 61 millones de personas viven con diabetes en las Américas, donde uno de cada dos adultos tiene sobrepeso u obesidad, uno de los factores de riesgo mayores de diabetes junto con la inactividad física. Además, se calcula que la diabetes está relacionada con más de medio millón de muertes cada año. Aunque esta proporción varía de acuerdo al acceso a los servicios de salud en cada país, se estima que una de cada tres personas con diabetes tipo 2 no lo saben y llegan al diagnóstico cuando ya están presentes las complicaciones de esta enfermedad.

La diabetes tipo 2, la forma más común de la enfermedad, puede prevenirse o retrasarse su aparición, pero la falta de conocimiento acerca de las medidas de prevención está generalizada (OMS, 2010). Entre estas medidas figuran alcanzar y mantener un peso corporal saludable; consumir una dieta saludable, que contenga entre tres y cinco raciones diarias de frutas y hortalizas y una cantidad reducida de azúcar y grasas saturadas; mantenerse activo físicamente, realizar al menos 30 minutos de actividad física moderada la mayoría de los días de la semana, y no fumar, ya que el consumo de tabaco aumenta el riesgo de enfermedades cardiovasculares.

La diabetes mal controlada aumenta el riesgo de muerte prematura, de enfermedad cardiaca y de accidente cerebrovascular, úlceras de los pies y, en última instancia, amputaciones de miembros inferiores; ceguera e insuficiencia renal. Además, las personas que padecen diabetes se encuentran en mayor peligro de presentar tuberculosis, especialmente aquellos con control glucémico deficiente. Según estudios, en los pacientes con diabetes el riesgo de muerte es al menos dos veces mayor que en las personas sin diabetes. A pesar del riesgo a la salud que está vinculado a la diabetes mal controlada, las personas que tienen diabetes pueden tener una vida saludable

si observan sistemáticamente las medidas de estilo de vida sana y siguen el tratamiento indicado.

La presión arterial alta, es decir igual o por encima de 140/90mmHg, aumenta el riesgo de tener un infarto al corazón, un accidente cerebrovascular e insuficiencia renal crónica. Distintos estudios estiman que la presión arterial alta contribuye a casi 9,4 millones de muertes al año en todo el mundo por enfermedades cardiovasculares. En las Américas, las enfermedades cardiovasculares causan 1,9 millones de muertes al año y son la principal causa de muerte en la mayoría de los países de la región. Conocer sus números es una responsabilidad personal, pero también es una responsabilidad profesional para médicos, enfermeras y trabajadores de la salud. La buena noticia es que la hipertensión es prevenible y tratable. En algunos países la prevención y el tratamiento de la presión arterial alta, así como de las causas, han llevado a una significativa reducción de las muertes por ataques cardíacos y accidentes cerebrovasculares.

Aunque la presión arterial alta afecta al 30% de la población adulta, una tercera parte desconoce tener esta condición. La hipertensión suele no dar síntomas, por esa razón es necesario que los adultos aprovechen cada oportunidad para medir su presión arterial. Además, los riesgos para la salud aumentan para quienes, además de tener la presión arterial elevada, fuman, son obesos o tienen diabetes. Sin embargo, se puede reducir el riesgo de hipertensión: consumiendo menos sal (en particular en los alimentos procesados), manteniendo una dieta balanceada y saludable, haciendo actividad física regularmente, evitando el uso de tabaco, evitando el consumo nocivo de alcohol.

Los países pueden trabajar en distintas intervenciones que ayuden a las personas a reducir el riesgo de hipertensión o a mantenerla bajo control. En promedio estamos consumiendo el doble de la cantidad de sal que necesitamos. Por eso la reducción de la sal en los alimentos es una de las

intervenciones claves para reducir la hipertensión. También es clave mantener una vida físicamente activa y consumir alimentos saludables. Disminuir el consumo de sal no depende únicamente del comportamiento individual. La mayor fuente de consumo de sal proviene hoy de los alimentos procesados y es justamente aquí donde deberíamos concentrar más esfuerzos, de la propia industria, pero también de los gobiernos y de toda la sociedad.

En los últimos años, los estudios sobre Enfermedades No Transmisibles se han incrementado, en dicho contexto a nivel nacional, desde el 2010 la Encuesta Demográfica y de Salud Familiar ha incorporado una batería de preguntas con la finalidad de captar y proveer información actualizada sobre la población afectada, para la evaluación y formulación de programas de salud orientados a disminuir o atenuar la prevalencia de las mismas (41% de la población de 60 años a más a nivel nacional tienen presión arterial alta, de acuerdo al ENDES 2013). Pero, la mayoría de estos estudios son descriptivos, ninguno ha analizado la relación estadística de asociación entre las Enfermedades No Transmisibles y las variables que la explican.

La presente investigación consiste en conocer cuáles son las variables y factores de salud asociados al comportamiento de las personas que influyen en la existencia de las enfermedades no transmisibles (Hipertensión Arterial y Diabetes Mellitus) en la población del departamento de Piura, que se pueden predecir con la utilización de un modelo Logístico Binario. Dada una significancia estadísticamente valida, y poder asumir aquellas variables y factores como características de riesgo, en la prevalencia de estas enfermedades.

La investigación se ha basado en el análisis de la Encuestas Demografía y Salud Familiar (ENDES) que efectúa el INEI, puesto que esta encuesta tiene información a nivel nacional y ofrecen una muestra muy amplia de casos en los cuales se puede obtener todos los beneficios del análisis estadístico. Las ENDES usada corresponden al departamento de Piura del año 2013.

1.2.2. Formulación del Problema

¿Cuáles son los factores de riesgo en el comportamiento de las personas asociadas a las enfermedades no transmisibles (Hipertensión Arterial y Diabetes Mellitus) en la población del departamento de Piura?

1.3. JUSTIFICACIÓN

Las enfermedades crónicas no transmisibles, afectan a todos los grupos de edad y constituyen un grupo heterogéneo de padecimientos como la diabetes e hipertensión arterial, entre otros; contribuyendo a un problema de salud pública por ser una causa de morbilidad, en el marco del proceso de envejecimiento de la población en nuestro país y por el modo de vida poco saludable (OMS, 2010, pág. i – 1). Según la Organización mundial de la Salud, las enfermedades crónicas son responsables del 60% de la mortalidad a nivel mundial y su prevalencia es mayor en la población adulta, situación en la que están inmersos los países en vías de desarrollo como el Perú.

Conocer sus números es una responsabilidad personal, pero también es una responsabilidad profesional para médicos, enfermeras y trabajadores de la salud. La buena noticia es que estas enfermedades son prevenibles y tratables. En algunos países la prevención y el tratamiento de estas enfermedades, así como de las causas, han llevado a una significativa reducción de las muertes por ataques cardíacos y accidentes cerebrovasculares. Por ello, la necesidad de la detección temprana, tratamiento y rehabilitación de los que padecen enfermedades no transmisibles y de planificar programas de acciones preventivas para esta población, que favorezcan su continuo desarrollo y mejore su calidad de vida.

A partir del año 2010, el Instituto Nacional de Estadística e Informática a través de la Encuesta Demográfica y de Salud Familiar incluyó una sección sobre traumatismo y enfermedades crónicas en el Cuestionario del Hogar. A partir de estos datos es posible elaborar indicadores sociodemográficos y de salud del

adulto mayor, con la finalidad de mejorar la evaluación y formulación de programas de salud orientados a reducir los factores de riesgos asociados a las enfermedades crónicas no transmisibles y fortalecer la atención sanitaria para los que las padecen. Pero, la mayoría de estos estudios son descriptivos, ninguno ha analizado la relación estadística de asociación entre las Enfermedades No Transmisibles y las variables que la explican.

La presente investigación consiste en conocer cuáles son las variables o factores de riesgo asociadas al comportamiento de las personas que influyen en la existencia de las enfermedades no transmisibles (Hipertensión Arterial y Diabetes Mellitus) en la población del departamento de Piura, que se pueden predecir con la utilización de un modelo Logístico Binario. Dada una confiabilidad estadística, y poder asumir aquellas variables o factores como características de riesgo, en la prevalencia de estas enfermedades.

La investigación se ha basado en el análisis de la Encuestas Demografía y Salud familiar (ENDES) que efectúa el INEI, puesto que esta encuesta tienen información a nivel nacional y ofrecen una muestra muy amplia de casos en los cuales se puede obtener todos los beneficios del análisis estadístico. Las ENDES usada corresponden al departamento de Piura del año 2013.

1.4. OBJETIVOS

1.4.1. Objetivo General

Identificar y establecer los factores de riesgo asociados al comportamiento de las personas que influyen en la existencia de las enfermedades no transmisibles (Hipertensión Arterial y Diabetes Mellitus) en la población del departamento de Piura, que se pueden predecir con la utilización de un modelo Logístico Binario.

1.4.2. Objetivos Específicos

1. Determinar a través de un modelo logístico binario los factores de salud individuales que influyen en la existencia de estas enfermedades no transmisibles (Diabetes Mellitus e Hipertensión), además de conocer las pruebas de bondad del ajuste del modelo y sus interpretaciones en cada caso.

2. Estimar a través del Odds ratio de cada factor significativo encontrado, la oportunidad de riesgo en desarrollar las enfermedades no transmisibles, en la población del departamento de Piura, que está expuesta a dichos factores de salud asociados a su comportamiento.

3. Encontrar las tablas de clasificación de ambos modelos de regresión logística binaria, además de conocer el porcentaje de aciertos global y la calidad predictiva a través del cálculo de la sensibilidad y especificidad para la variable dependiente.

4. Determinar a través de la curva ROC el valor de mejor sensibilidad y especificidad para ambos modelos y cuál es el punto de corte óptimo, para mejorar su calidad predictiva.

1.5. FORMULACIÓN DE LA HIPÓTESIS:

Existen factores de riesgo asociados al comportamiento de las personas identificadas con el modelo logístico binario que influyen en la existencia de las enfermedades no transmisibles (Hipertensión Arterial y Diabetes Mellitus) en la población del departamento de Piura.

1.6. LIMITACIONES

Esta investigación presenta las siguientes limitaciones:

- Por temas de seguridad de la información y políticas internas de la institución donde laboro, no es posible, mostrar ni acceder a toda la información de las personas encuestadas como una exposición pública dado que por carácter de privacidad regulado por la ley, esta se maneja de forma confidencial según el D. L. N° 604 –secreto estadístico.

- No se cuenta con modelos anteriores que permitan tener una base para generar un modelo muy complejo. Solo se tienen trabajos de investigación con estas enfermedades de tipo descriptivo, no se pueden establecer comparaciones puesto que usar modelos estadísticos no se relaciona con evaluaciones de esa naturaleza.

1.7. DELIMITACIONES

Teniendo en cuenta los criterios de la investigación científica, esta investigación se ha delimitado de la siguiente manera:

- La información con la que se cuenta ha sido obtenida durante todo un año de trabajo y que comprende a todo el proceso del estudio de salud en referencia para el año 2013, y que se mantiene guardada en las bases de datos de la institución y que han sido proporcionadas para los fines de esta investigación.

- En esta investigación no se cuenta con factores de carácter personal, social, económicos, cultural, para las personas implicadas en nuestras encuestas, es por ello que no se tiene ese tipo de información en la base de datos.

- Las encuestas han sido realizadas por personal del INEI, que como en mi caso desarrollamos esta recepción de información como parte de nuestra labor profesional y que se ha llevado a cabo en todo el departamento de Piura.

II. MARCO TEÓRICO

2.1. ANTECEDENTES GENERALES

A Nivel Internacional

Las enfermedades no transmisibles (ENT) afectan ya desproporcionadamente a los países de ingresos bajos y medios, donde se registran casi el 80% de las muertes por ENT (29 millones). Son la principal causa de mortalidad en todas las regiones excepto en África, pero según las estimaciones actuales en 2020 los mayores incrementos de la mortalidad por ENT corresponderán a ese continente. En los países africanos, se prevé que las defunciones por ENT superarán la suma de las causadas por las enfermedades transmisibles y nutricionales y por la morbilidad materna y perinatal como causa más frecuente de muerte en 2030.

Más de 9 millones de las muertes atribuidas a las enfermedades no transmisibles se producen en personas menores de 60 años de edad; el 90% de estas muertes prematuras ocurren en países de ingresos bajos y medianos. Las enfermedades cardiovasculares constituyen la mayoría de las defunciones por ENT (17,3 millones cada año), seguidas del cáncer (7,6 millones), las enfermedades respiratorias (4,2 millones), y la diabetes (1,3 millones). Estos cuatro grupos de enfermedades son responsables de alrededor del 80% de las muertes por ENT. Además, comparten cuatro factores de riesgo: el consumo de tabaco, la inactividad física, el uso nocivo del alcohol y las dietas malsanas.

La hipertensión es el principal factor de riesgo de muerte en el mundo y afecta tanto a hombres como a mujeres. Al menos tres de cada diez adultos en la región de las Américas tiene presión arterial alta o hipertensión, el principal factor de riesgo para enfermedades cardiovasculares y muertes en todo el mundo. Se estima que la hipertensión afecta a casi 1000 millones de personas en todo el mundo (OPS-OMS, 2013). Aunque la presión arterial por encima de 140/90mmHg afecta al 30% de la población adulta, una tercera parte desconoce su enfermedad. Tres de

cada Diez personas que se está tratando por hipertensión no consigue mantener su presión arterial por debajo del límite de 140/90. La información disponible en algunos países, como en los EEUU, revela que mientras la hipertensión arterial es más frecuente en hombres, a partir de edades superiores a los 65 años existe una elevada proporción de mujeres con hipertensión arterial. También se observa una proporción elevada de personas afro-descendientes con hipertensión arterial, que afecta tanto a hombres como a mujeres.

Tener la presión arterial controlada es clave. Estudios recientes muestran que las tasas de control de la hipertensión o sea, de quienes consiguen mantener la presión arterial por debajo de 140 y 90, suelen ser bajas en América Latina, en tanto van entre 12% y 41%. Aun así, hay países como Canadá, Cuba y Estados Unidos que han progresado en sus tasas de control, que en la actualidad están en más del 50%.

Además de las iniciativas de prevención que muchos países de las Américas han desarrollado, también es importante que los proveedores de salud aseguren una detección temprana y un adecuado tratamiento de la hipertensión, afirmó Pedro Ordoñez, asesor en Enfermedades no Transmisibles de la OPS/OMS. Las personas que son diagnosticadas con hipertensión pueden ser tratadas y controladas a largo plazo, lo que mejora significativamente su probabilidad de tener vida larga, saludable y productiva.

Las autoridades sanitarias de las Américas aprobaron en 2012 una estrategia para la prevención y el control de las enfermedades no transmisibles, que tiene por meta reducir en un 25% la mortalidad prematura por enfermedades cardiovasculares, cáncer, diabetes y enfermedades respiratorias crónicas para 2025. De alcanzar esta meta, fijada por la Asamblea Mundial de la Salud ese año, se estima que se salvarán tres millones de vidas en la región. Con esta estrategia, los países de las Américas se han comprometido, priorizar las enfermedades no transmisibles e incorporarlas a las políticas de salud y de desarrollo; establecer mecanismos multisectoriales para promover el diálogo y las asociaciones entre

gobiernos y sectores no gubernamentales; y fortalecer las medidas que tiendan a reducir los factores de riesgo y mejorar la cobertura de la atención.

En las Américas, la obesidad y la diabetes están afectando a la población con tasas cada vez más elevadas. Las encuestas nacionales demuestran que la prevalencia de la obesidad está aumentando en todos los grupos de edad. Entre el 7% y 12% de los niños menores de 5 años y una quinta parte de los adolescentes son obesos, mientras que en los adultos se aproximan al 60%. La obesidad es el principal factor de riesgo de la diabetes. Se prevé que el número de personas que sufren diabetes en América Latina se incremente en más de un 50% en los próximos 15 años pasando de 13,3 millones en el 2000 a 32,9 millones en el 2030. La diabetes y la obesidad, afectan desproporcionadamente a los sectores pobres y de nivel cultural más bajo.

La diabetes mellitus es una enfermedad metabólica crónica, caracterizada por altos índices de glucosa en la sangre (hiperglucemia) y asociada a una deficiencia absoluta o relativa en la secreción o acción de la insulina. Hay tres formas principales de diabetes: la diabetes de tipo 1, la de tipo 2 y la diabetes gestacional. La diabetes de tipo 2 es la más común; representa aproximadamente entre el 85 y 90% de los casos, y se relaciona con factores de riesgo modificables como la obesidad o el sobrepeso, la inactividad física y los regímenes alimentarios hipercalóricos y de bajo valor nutritivo.

La hiperglucemia intermedia, a menudo llamada prediabetes, es un componente del síndrome metabólico el cual se caracteriza por la presencia de prediabetes junto con algún factor de riesgo de las enfermedades cardiovasculares como por ejemplo: la hipertensión arterial, la obesidad en el segmento superior del cuerpo o la dislipidemia. Cálculos recientes revelan que, en los países latinoamericanos y del Caribe las tasas más elevadas de prevalencia de la diabetes corresponden a Belice (12,4%) y México (10,7%). Managua, Ciudad de Guatemala y Bogotá mantienen tasas de alrededor del 8 al 10%. Estados Unidos representa una prevalencia de alrededor del 9,3%, llegando a prácticamente el 16%, en la frontera

mexicoestadounidense. La carga que representa la diabetes para las personas y la sociedad se relaciona principalmente con un aumento de la discapacidad y la mortalidad prematura causada por las complicaciones de esta enfermedad.

En un estudio clínico realizado en seis países latinoamericanos se halló que, la frecuencia de complicaciones crónicas en personas que han padecido diabetes durante más de veinte años son del 48% para las retinopatías, 6,7% para la ceguera, 42% para las neuropatías, 1,5% para el daño renal, 6,7% para el infarto de miocardio, 3,3% para los accidentes cerebrovasculares y 7,3% para las amputaciones de los miembros inferiores. Si bien la diabetes y sus complicaciones son en gran medida prevenibles, con mucha frecuencia se carece de conocimientos acerca de las medidas de prevención y no hay acceso a servicios de atención adecuados.

La comunidad internacional ha reconocido el problema de las enfermedades crónicas y ha establecido la forma de combatirlas por medio de la Estrategia Mundial de la OMS para la Prevención y Control de las Enfermedades Crónicas (WHA53.17, 2000), el Convenio Marco para el Control del Tabaco (WHA56.1, 2003), la Estrategia Mundial sobre Régimen Alimentario, Actividad Física y Salud (WHA57.17, 2004) y, últimamente, la ya mencionada Estrategia Regional para las enfermedades crónicas (CD47/17, Rev.1).

A Nivel Nacional

Las enfermedades cardiovasculares continúan siendo la principal causa de muerte en el mundo; así como, de morbilidad y pérdida de calidad de vida relacionada con la salud. En el caso del Perú Las enfermedades isquémicas del corazón y las enfermedades cerebrovasculares se constituyen como segunda y tercera causa de mortalidad en el adulto mayor, respectivamente (Análisis de la Situación de Salud del Perú, 2010. Pág. 51).

En nuestro país, las Enfermedades Crónicas o No Transmisibles representan el 58.5% de las enfermedades con mayor incidencia, al mismo tiempo son estas enfermedades las que producen mayor discapacidad. La prevalencia el año 2011 de personas con Hipertensión Arterial fue de 198,925 (17.9%) con una mortalidad de 21.2 x 100 mil habitantes; asimismo las personas con Diabetes Mellitus tuvieron una prevalencia de 104,227 (3.6%) con una mortalidad de 18,9 x 100 mil habitantes.

Las enfermedades crónicas o no transmisibles, afectan a todos los grupos de edad y constituyen un grupo heterogéneo de padecimientos como la diabetes e hipertensión arterial, entre otros; contribuyendo a un problema de salud pública por ser una causa de morbilidad, en el marco del proceso de envejecimiento de la población en nuestro país y por el modo de vida poco saludable (OMS-Informe sobre la Situación Mundial de las Enfermedades No Transmisibles, 2010, pág. i – 1).

Este crecimiento de la población adulta mayor y el aumento de la expectativa de vida de los peruanos que han pasado de 45 años en 1960 a 71.8 años en 2011, está produciendo un incremento significativo de enfermedades degenerativas propias de dichas edades como el cáncer, hipertensión arterial, diabetes Mellitus, arterioesclerosis, hiperlipidemias, obesidad y enfermedades mentales asociadas con carencias secundarias.

Este proceso denominado de transición epidemiológica, hace que las enfermedades infectocontagiosas están siendo desplazadas por las enfermedades crónicas, requiere de una correcta investigación para valorar el impacto que éstas enfermedades están produciendo en nuestra población y adecuar los servicios de salud a éstas nuevas demandas, así como la programación de líneas educativas para la población y los profesionales de la salud.

Entre las principales causas de mortalidad por enfermedades crónicas no transmisibles están las enfermedades cerebrovasculares (31,4%), la diabetes

mellitus (20,4%), la enfermedad hipertensiva (17,1%), Los resultados de la Encuesta Demográfica y de Salud Familiar 2012 mostraron que, un 29,7% de la población adulta mayor declaró haber sido informada en algún momento por un médico o profesional de la salud que padece de presión alta o hipertensión arterial; un 70,1% seguían tratamiento antihipertensivo farmacológico y el 29,9% a ningún tratamiento. Al comparar con el año 2011, se aprecia un incrementó de 1,6 puntos porcentuales en los adultos mayores con presión alta y de 3,7 puntos porcentuales entre los que no siguieron tratamiento.

Según la ENDES, muestran que el 8,7% de los adultos mayores declararon haber sido informados que tenían diabetes y que el 78,7% recibieron tratamiento médico para tener mejor calidad de vida; sin embargo, es preocupante que el 21,3% de adultos mayores no hayan buscado tratamiento por un profesional de la salud exponiéndose a diversas complicaciones asociadas a la enfermedad. Con relación al año 2011, la proporción de adultos mayores con diabetes (7,8%) se incrementó en 1,1 puntos porcentuales. Según ENDES, la población de 15 y más años de edad con presión arterial medida, se encontró un 16,6% con hipertensión arterial; siendo los hombres más afectados (21,5%) que las mujeres (12,3%). La prevalencia de hipertensión arterial es mayor en la Costa sin Lima Metropolitana (20,2%) seguido por Lima Metropolitana (18,7%); en tanto, la menor prevalencia se registró en la Selva (11,9%) y en la Sierra (13,3%) (Perú: Enfermedades No Transmisibles y Transmisibles, 2013).

El consumo de tabaco aumenta el riesgo de las enfermedades no transmisibles. Este comportamiento afecta a un 20,6% de la población de 15 y más años de edad que fumaron un cigarrillo (de manufactura industrial o artesanal). Según sexo, el consumo de cigarrillo es más recurrente en los hombres (34,6%) que las mujeres (8,3%). De acuerdo con la región natural de residencia, se encontró en Lima Metropolitana (26,6%) y la Selva (21,3%) mayores porcentajes de consumo de cigarrillos en la población de 15 y más años de edad; no obstante, entre los menores porcentajes se ubican la Sierra (16,7%) y Costa sin Lima Metropolitana (16,9%), (Perú: Enfermedades No Transmisibles y Transmisibles, 2013). La

ENDES, manifestó que el 88,0% de las personas de 15 y más años de edad declararon que han consumido bebida alcohólica, alguna vez en su vida. Siendo mayor este porcentaje en los hombres, 92,9% que en las mujeres, 83,8%. (Perú: Enfermedades No Transmisibles y Transmisibles, 2013).

La baja ingesta de frutas y verduras contribuyen con el desarrollo de enfermedades cardiovasculares, cáncer, diabetes o la obesidad. Es considerado entre los principales factores de riesgos comportamentales, que prevalece en áreas rurales, población con bajos ingresos económicos y bajo nivel educativo (OMS-Informe sobre la Situación Mundial de las Enfermedades No Transmisibles, 2010.Pág 5). El consumo de verduras o vegetales es un componente importante para una dieta saludable, y el consumo diario podría prevenir enfermedades cardiovasculares y algunos cánceres, por ello es necesario contar con cifras estadísticas que permitan monitorear la ingesta de verduras (excluidas las papas y otros tubérculos feculentos).

El síndrome metabólico está compuesto por una serie de anormalidades que incluyen obesidad abdominal, anormalidades del metabolismo de la glucosa, hipertensión, y dislipidemia acompañado de un estado pro trombotico y pro inflamatorio el cual lleva en el tiempo al desarrollo de diabetes mellitus II, así como enfermedad vascular (enfermedad coronaria y enfermedad vascular cerebral).

A Nivel Local

Los antecedentes de tipo regional y local respecto a las enfermedades que son parte de esta investigación, no se pueden describir, puesto que no existe ninguna investigación similar en estos dos contextos. Por lo tanto, lo que se puede incluir una apreciación subjetiva con respecto a las personas que padecen este tipo de enfermedades en función al conocimiento que tengo por ser parte de las actividades de trabajo que realiza la ENDES, desde el año 2013. El departamento de **Piura ocupa el segundo lugar en el Perú** en cantidad de casos de **diabetes** y los casos de **hipertensión arterial**, colesterol alto y obesidad y estas aumentan a

un ritmo acelerado, la cifra es alarmante si consideramos que antes los pacientes con estas enfermedades tenían entre 70 y 80 años de edad, pero ahora aparecen personas desde 35 o 40 años, que ya tienen un infarto. Estas complicaciones se dan de acuerdo al estilo de vida de la población en la actualidad.

Cuando se entrevistó, informantes manifiestan que a pesar que fueron informados por un personal de salud que tiene la enfermedad, solo unos cuantos reciben tratamiento médico para tener mejor calidad de vida, sin embargo es preocupante que la población que no busca consejo o tratamiento está expuesta a diversas complicaciones asociadas a la enfermedad.

2.2. ENFERMEDADES NO TRANSMISIBLES O CRÓNICAS

Las enfermedades crónicas son aquellas enfermedades de larga duración, degenerativas, no transmisibles mediante el contacto personal y generalmente, interfieren con la capacidad del cuerpo para funcionar de manera óptima. Estas enfermedades no son curables pero pueden ser prevenidas o evitadas con intervenciones para reducir los factores de riesgo que conducen al desarrollo de las mismas por ejemplo, evitando o minimizando el consumo de tabaco, del alcohol, aumentando la actividad física y consumiendo una dieta balanceada.

Dentro de las enfermedades crónicas, son de citar: Diabetes, colesterol y/o triglicéridos altos, enfermedades renales, asma, cáncer, enfermedades cardiovasculares, hipertensión arterial, entre otras; siendo las principales enfermedades no transmisibles la Hipertensión Arterial y Diabetes Mellitus. Según la Organización mundial de la Salud, las enfermedades crónicas son responsables del 60% de la mortalidad a nivel mundial y su prevalencia es mayor en la población adulta, situación en la que están inmersos los países en vías de desarrollo como el Perú.

Con tal motivo, el Ministerio de Salud ha implementado la **Estrategia Sanitaria Nacional de Prevención y Control de Daños No Transmisibles**, a fin de mediar una acción preventiva que modifique el nivel de los factores de riesgo de estas enfermedades entre la población de nuestro país y disminuir la morbilidad y mortalidad causadas por estas enfermedades.

En los últimos años, los estudios sobre Enfermedades No Transmisibles se han incrementado, en dicho contexto, desde el 2010 la Encuesta Demográfica y de Salud Familiar - ENDES ha incorporado una batería de preguntas con la finalidad de captar y proveer información actualizada sobre la población afectada, para la evaluación y formulación de programas de salud orientados a disminuir o atenuar la prevalencia de las mismas (41% de la población de 60 años a más a nivel nacional tienen presión arterial alta, de acuerdo al ENDES 2013). Pero, la mayoría de estos estudios son descriptivos, ninguno ha analizado la relación estadística de asociación entre las Enfermedades No Transmisibles y las variables que la explican.

Las enfermedades materia de investigación son:

2.2.1. Diabetes o Azúcar Alta en la Sangre

Es el resultado de un desorden en el metabolismo de los alimentos en el cuerpo humano, la misma que se caracteriza por una elevada concentración de glucosa en la sangre; ya sea por falta de la hormona llamada insulina, segregada por el páncreas, que posibilita su absorción por las células del cuerpo humano; o, por que las células del cuerpo no responden al estímulo de la insulina generada por el páncreas (Manual de la entrevistadora - ENDES, 2013, pág.107).

La diabetes mellitus es una enfermedad que se clasifica según su causa. La Organización Mundial de la Salud y la Asociación Latinoamérica de Diabetes Mellitus consideran:

a. **Diabetes Tipo 1:** en este tipo (por problema autoinmunes), el cuerpo destruye las células beta (β) del páncreas (el tejido productor de insulina en el cuerpo).

b. **Diabetes Tipo 2:** en este tipo (que tiene diferente causa), el cuerpo hace resistencia a la insulina y posteriormente se produce una disminución de su producción.

c. **Diabetes Gestacional:** en este tipo (ocurre en el embarazo), la madre gestante llega a tener niveles muy elevados de glucosa, pero solo durante el embarazo.

d. **Diabetes de Otro Tipo:** ocurre por una diversidad de motivos (enfermedades del páncreas, hormonales, genéticas, uso de medicamentos).

La Diabetes Tipo 2, la de mayor ocurrencia, se produce por el deterioro progresivo de las células β del páncreas, a consecuencia de la resistencia a la insulina: el cuerpo disminuye el uso de la glucosa (azúcar) por los tejidos (musculo, grasa, hígado), produciendo niveles altos de glucosa en la sangre, que acelera la muerte de las células β del páncreas, y aumenta el nivel de ácidos grasos libres (lo que reduce el efecto de la insulina en los tejidos).

El riesgo aumenta con la edad, la obesidad y la inactividad física. El nivel de glucosa en la sangre se eleva gradualmente y sus síntomas (orinar mucho, tomar mucha agua, y pérdida de peso) pueden pasar desapercibidos. Con el tiempo el nivel elevado de glucosa produce daño, falta de funciones y de varios órganos vitales (ojo, riñones, sistema nervioso, corazón y vasos sanguíneos).

La prediabetes es un estado anormal (glucosa alta en ayunas, y nivel elevado de glucosa en la sangre después de 2 horas de comer) que ocurre por lo menos 10 años antes de desarrollarse la diabetes. Las células β del páncreas entran en disfunción a causa de la resistencia a la insulina. Se considera diagnóstico

de diabetes mellitus (es decir, un caso confirmado) cuando la persona presente cualquiera de los siguientes criterios:

Criterio	Característica
Síntomas y Glicemia casual (al azar)	**Poliuria** (aumento de la cantidad de orina), **polidipsia** (aumento anormal de la sed) y **perdida** inexplicable de **peso** además de glicemia casual de al menos 200 mg/dl
Glicemia en ayunas (al menos 8 horas)	Resultado de al menos **126mg/dl en dos oportunidades** (entre no más de 72 horas entre ellas).
Tolerancia a la glucosa	Resultado de al menos **200mg/dl** dos horas después de una prueba de una carga de **75g de glucosa.**

El objetivo inmediato del tratamiento de la diabetes mellitus es lograr el control glucémico (debajo de los criterios diagnósticos en ayunas y/o de tolerancia), así como el control de síntomas. Las medidas generales y preventivas incluyen el tratamiento dietético con un nutricionista, y la consejería nutricional, cuyas pautas son: evitar consumir azucares refinados; reducir el consumo de grasa; no mezclar harinas; consumir 8 vasos de agua natural; fraccionar de comidas (5 al día); evitar el consumo de tabaco.

La terapia farmacológica (medicamentos) incluye para todos los esquemas posibles de tratamiento, iniciar con un medicamento (llamado metformina) y según el nivel de glicemia añadir una sulfonilurea o insulina basal.

2.2.2. Presión Alta o Hipertensión Arterial

Es la elevación persistente de la presión arterial por encima de los límites considerados como normales. La presión arterial es la fuerza con que la sangre empuja las paredes de los vasos sanguíneos, la misma que en el registro medico se escribe con dos números, por ejemplo, 110/80 mm Hg (Milimetros de mercurio), (Manual de la entrevistadora - ENDES, 2013, pág.109).

El primer número alude a la presión sistólica, la cual se genera cuando el corazón se contrae bombeando la sangre hacia los vasos sanguíneos. El segundo número corresponde a la presión diastólica, la que genera cuando el corazón se llena de sangre mientras se relaja entre latidos. La presión alta o hipertensión arterial se define como la Presión Arterial Sistólica de 140mmHg o más, o una Presión Arterial Diastólica de 90mmHg o más. La presión es una enfermedad silenciosa (asintomático) y fácil de detectar, generalmente no puede curarse pero si controlarse. De no tratarse, la presión alta presente peligros y aumenta el riesgo de ataques al corazón y/o ataques al cerebro.

La enfermedad hipertensiva es un síndrome, cuyo componente indispensable es la elevación anormal de la presión arterial sistólica y/o diastólica. Para la población adulta, se consideran cifras patológicas a la presión arterial sistólica mayor o igual a 140mmHg y a la diastólica mayor o igual a 90mmHg y es el responsable del 29.2% de muertes (la mayoría prevenibles), del 30% de los casos de insuficiencia renal crónica y es el factor de riesgo más importante de los accidentes cerebrovasculares (75%).

En los vasos sanguíneos aumenta el tejido interno (endotelio), aumenta su resistencia, y aumentan las sustancias en la cavidad del vaso. Todo ello explica el crecimiento anormal del corazón y el desarrollo de complicaciones de los vasos sanguíneos (aterosclerosis), del corazón (insuficiencia cardiaca, isquemia miocárdica y arritmias), del cerebro (hemorragia, isquemia cerebral y encefalopatía) y del riñón (insuficiencia renal). Generalmente no tiene síntomas (en el 70% a 80% de los casos).

Se clasifica el nivel de hipertensión arterial de acuerdo a los límites que figuran en la tabla siguiente. Cuando la presión sistólica o diastólica se sitúa en categorías diferentes, la presión mayor debe ser utilizada para la clasificación o estadio.

Categoría	Sistólica (mmHg)	Diastólica (mmHg)
Normal	<120	<80
Pre- Hipertensión	120-139 y/o	80-89
Hipertensión		
Estadio 1	140-159 y/o	90-99
Estadio 2	≥160-179 y/o	≥100-109

El objetivo del tratamiento de la hipertensión arterial es alcanzar niveles menores a 140/90 (presión sistólica/ presión diastólica), lo que permitirá la reducción de las complicaciones cardiovasculares (en personas con diabetes, síndrome metabólico o con enfermedad renal).

Las medidas generales y preventivas incluyen la modificación del estilo de vida: mantener o alcanzar el peso corporal ideal; realizar actividad física; reducir el consumo de sal; reducir el consumo de grasa saturadas; evitar el consumo de alcohol; evitar el consumo de tabaco. La terapia farmacológica (medicamentos) incluye, para los esquemas posibles, iniciar con un medicamento (llamado IECA, inhibidor de la enzima convertidora de angiotensina), y según el nivel del logro del objetivo, añadir otro medicamento (calcioantagonista).

2.3. FACTORES DE RIESGO ASOCIADOS A LAS ENFERMEDADES CRÓNICAS

La prevalencia de las enfermedades crónicas o no transmisibles es propia de la población adulta mayor que se sustenta en diversos factores de riesgo: actividad física, hábitos alimenticios, consumo de cigarrillo o tabaco y de alcohol. La evidencia demuestra que la modificación de hábitos no saludables y el control de los factores de riesgo pueden, en la mayoría de los casos, evitar las manifestaciones clínicas de algunas enfermedades crónicas e impedir complicaciones que, sin control, pueden causar discapacidades que tienden a disminuir la calidad de vida de las personas mayores (Manual sobre indicadores de calidad de vida en la vejez, Chile, 2006. Pág. 82). Por ello, identificar estos

factores es importante para la prevención primaria en beneficio de la salud del adulto mayor.

Un factor de riesgo es toda circunstancia o situación que aumenta las posibilidades de una persona de contraer una enfermedad. Son características y atributos que se presentan asociados a la enfermedad. Los factores de riesgo no son necesariamente las causas de las enfermedades. Se han identificado factores de riesgo asociados a las enfermedades crónicas, como tabaquismo, consumo excesivo e inapropiado de alcohol, inactividad física, obesidad, perfil lipídico alterado y dieta inadecuada. Muchos de estos factores de riesgos son comunes a varias de estas enfermedades. Esto refleja que existe una oportunidad de daño que aún no se ha manifestado clínicamente.

Podemos suponer que en muchos individuos la historia natural de las enfermedades de este grupo está en etapas tempranas, subclínicas, e incluso en gente joven, en las cuales el daño se está recién iniciando y puede ser aun reversibles. En estos grupos, una oportunidad y eficiente intervención impediría o retardaría el curso inexorable de las enfermedades crónicas no transmisibles. Por tanto, de no mediar una acción preventiva que modifique estos factores de riesgo, el diagnóstico y tratamiento de las Enfermedades o Daños No Transmisibles en nuestro país alcanzará en los próximos años cifras realmente epidémicas.

Los Factores de riesgo materia de investigación son: Actividad física, dieta alimentaria, tabaquismo y consumo de alcohol.

2.3.1. Actividad Física

La capacidad de desempeñarse normalmente en actividades diarias, el esfuerzo físico que les demanda alguna tarea, la práctica de algún ejercicio físico y el número de horas que pasa viendo televisión, se ve afectada por la edad; la cual puede, estar a su vez relacionada con la presencia de

enfermedad, padecimientos crónicos o lesiones que afecten las habilidades físicas o mentales del adulto mayor.

La actividad física es cualquier movimiento voluntario producido por la concentración musculo esquelético, que tiene como resultado un gasto energético que se añade al metabolismo basal. La actividad física se presenta en todas las actividades cotidianas como trabajar, caminar, realizar quehaceres domésticos, en cambio, el ejercicio es planeado, estructurado y repetitivo con un mayor o menor consumo de energía, su finalidad es producir un mejor funcionamiento del propio organismo. Para la OMS, la actividad física es un pilar de la prevención frente a las causas de muerte actuales, ya que mantiene la salud y reduce los riesgos de mortalidad por todas las causas.

La actividad física puede ser clasificada de varias maneras, incluyendo tipo, intensidad y propósito. La actividad física es un medio fundamental para mejorar la salud física y mental. Reduce alrededor del 50% el riesgo de muchos trastornos relacionados con inactividad (como las enfermedades del corazón y diabetes tipo II), reduce sustancialmente el riesgo de la hipertensión y algunas formas de cáncer, y también la disminución del estrés, ansiedad, depresión y soledad. La actividad física regular ayuda a proteger contra el aumento de peso no saludable.

La evidencia científica muestra que el sedentarismo tiene un impacto negativo en la salud y es un importante factor contribuyente a un amplio rango de enfermedades crónicas como enfermedad coronaria, accidente vascular o hipertensión arterial. La prevalencia de estilos de vida sedentarios sigue aumentando, por lo que son necesarias intervenciones de promoción de la actividad física que permitan alcanzar el objetivo de acumular al menos 30 minutos de actividad física de intensidad moderada en casi todos, o todos, los días de la semana.

2.3.2. Dieta alimentaria

La manera de alimentarse de cada persona es un reflejo no sólo de los hábitos aprendidos, sino también de la propia forma de pensar. Las personas de 60 y más años de edad tienen bien establecidos sus hábitos de comida que se han consolidado con el paso de los años. No obstante, el inadecuado hábito alimenticio aumenta el riesgo de padecer alguna enfermedad crónica no transmisible: cardiopatías, enfermedades cardiovasculares, diabetes, hipertensión arterial y algún tipo de cáncer.

La dieta alimentaria es la cantidad de principios alimentarios que deben ingerirse diariamente para poder satisfacer las necesidades del organismo. Es por ello, que una correcta dieta alimentaria debe asegurar al organismo: la energía necesaria para su normal funcionamiento (alimentos energéticos); las sustancias que intervienen en la formación, crecimiento y mantenimiento de los tejidos (alimentos plásticos o tisulares) y la presencia de cantidades mínimas de sustancias que regulan su funcionamiento (vitaminas, sales minerales).

La dieta alimentaria de un individuo debe proporcionar una suficiente cantidad de calorías adecuadas al sexo, a la edad, al tipo de actividad que desarrolla. Cuando el hombre realiza una actividad intensa, la necesidad energética puede llegar a valores muy superiores. En cambio, la actividad mental o sedentaria no exige complementos energéticos.

La dieta alimentaria es suficiente cuando la cantidad de alimentos ingerida cubre las necesidades para mantener la salud. Además debe haber una alimentación variada mediante combinaciones adecuadas de productos de origen vegetal y animal. El organismo necesita ingerir todas las sustancias que lo forman, es decir, debe incorporar los prótidos, glúcidos, lípidos, agua, sales minerales y vitaminas. La falta o reducción de uno de ellos produce un

régimen deficiente o de carencia. Una alimentación inadecuada, rica en sal y grasas saturadas aumenta la presión arterial.

Consumo de frutas y verduras.

La baja ingesta de frutas y verduras contribuyen con el desarrollo de enfermedades cardiovasculares, cáncer, diabetes o la obesidad. Es considerado entre los principales factores de riesgos comportamentales (OMS-Informe sobre la Situación Mundial de las Enfermedades No Transmisibles 2010, Pág. 5), que prevalece en áreas rurales, población con bajos ingresos económicos y bajo nivel educativo.

La dieta, considerada como el consumo generoso de frutas y verduras, disminuye el riesgo de enfermedades cardiovasculares, accidente cerebro vascular, cáncer de estómago y cáncer de colon y de recto. Este efecto se da por las altas ingestas de potasio y ácido fólico, así como de fibra alimentaria. La mayor parte de las fibras reducen el colesterol en la sangre y el solo hecho de consumir frutas y verduras frescas reduce ligeramente la presión arterial sistólica y diastólica, pero con una reducción considerable de riesgo de enfermedad cerebro vascular. Además, un alto consumo de frutas y verduras limitara el consumo excesivo de calorías. La meta de la OMS es que se consuma al menos 400g de fruta y/o verduras diariamente, traducido como 5 porciones diarias.

2.3.3. Tabaquismo

Si bien el rol del tabaco como factor causal de cáncer de pulmón es bastante conocido, este también está asociado al desarrollo de enfermedad vascular y por ende constituye un factor de riesgo para el desarrollo de enfermedades isquémicas. Actualmente se calcula que las enfermedades cardiovasculares son responsables de casi la mitad de los decesos vinculados al consumo de tabaco en países desarrollados así como más de un cuarto de estos en países en vías de desarrollo.

El consumo de tabaco aumenta el riesgo de las enfermedades no transmisibles. Este comportamiento afecta a un 20,6% de la población de 15 y más años de edad que fumaron, en los últimos 12 meses del año 2013, al menos un cigarrillo (de manufactura industrial o artesanal) (Perú: Enfermedades No Transmisibles y Transmisibles, 2013).

El tabaco es uno de los peores enemigos del sistema circulatorio, ya que aumenta la presión arterial y la frecuencia cardiaca. Además de producir numerosas enfermedades vasculares impide que los tratamientos o las sustancias que podrían ayudar al hipertenso sean absorbidos por el organismo. Las personas fumadoras no se benefician tanto de los medicamentos contra la hipertensión como aquellos que no fuman.

Hábito de fumar

El hábito de fumar causa la muerte de 6 millones de personas cada año, causa el 71% del cáncer pulmonar, el 42% de enfermedades respiratorias cónicas y el 10% de las enfermedades cardiovasculares. La mortalidad de los fumadores se asocia a la cantidad de cigarrillos fumados, al tiempo de consumo, a la profundidad de la inhalación y al contenido de alquitrán y nicotina fumado.

Las sustancias químicas contenidas en el tabaco son precursoras de más de 4000 sustancias en el humo de la combustión, además de agentes químicos del grupo I de carcinógenos humanos (benceno, cadmio, astato, niquel, cromo, 2 naftilamino, cloro vinil, aminobifenil y berilio) y la nicotina es el responsable de la adicción al tabaco, cuya inhalación profunda (1 a 2 mg por cigarrillo) pasa al cerebro muy rápidamente (entre 9 a 10 segundos) y después a todos los tejidos corporales. Estas producen descargas de adrenalina, de glucosa, aumento de presión arterial, ritmo cardiaco y metabolismo basal, y formación de trombos, aumentado de colesterol, aumento de riesgo de gastritis y ulceras, infecciones pulmonares y reduce la formación de neuronas.

Los efectos más importantes son el cáncer (de pulmón, laringe, esófago, cavidad oral, vejiga, riñón, etc.), la enfermedad cardiovascular (isquemia coronaria, infarto del miocardio, accidente cerebrovascular, aterosclerosis), la enfermedad respiratoria (bronquitis, asma, etc.), el mayor riesgo de depresión, adicción al alcohol, enfermedades psiquiátricas, desnutrición, consecuencias en mujeres (infertilidad, retraso de la concepción, adelanto de la menopausia , incremento de osteoporosis), consecuencia en el embarazo (placenta previa, parto prematuro, retraso de desarrollo cerebral malnutrición fetal, bajo peso al nacer) y post embarazo (retraso del crecimiento postnatal y desarrollo cognitivo), entre otros efectos.

El inicio del consumo se da a muy temprana edad (a partir de los 12 años). Los adolescentes empiezan a fumar principalmente por influencias sociales (sus padres, compañeros, amigos, hermanos y medios de comunicación). El fumador es la persona que consume tabaco (industrial o artesanal); sin embargo, para esta investigación consideramos al cigarrillo como producto representativo derivado del tabaco.

2.3.4. Consumo de alcohol

Muchos especialistas afirman que un uso moderado de alcohol favorece la circulación, puede disminuir la hipertensión y previene muchas enfermedades del corazón y del sistema circulatorio. Sin embargo, el consumo no moderado del alcohol conlleva al aumento de la presión arterial y la posibilidad de alcoholismo. Una ingesta superior a 40 gramos de alcohol puede producir el aumento de la presión arterial.

Algunas veces hay que restringir aún más el consumo de alcohol, sobre todo si la hipertensión se asocia a enfermedades metabólicas como diabetes o el aumento de ciertas grasas en la sangre (triglicéridos). El consumo de alcohol excesivo ocasiona la muerte de 2.3 millones de personas cada año, el 3.8% de todas las muertes en el mundo. Más de la mitad de estas muertes es por

enfermedades no transmisibles como las cardiovasculares (diabetes, enfermedad hipertensiva, cardiopatía isquémica, accidente cerebro vascular) cánceres (de colon, de recto, de mama, etc.), cirrosis, entre otras.

El consumo de alcohol aumenta el riesgo de consecuencias adversas a la salud si el hábito persiste (siendo de 20 a 40g diarios en mujeres y 40 a 60g diarios en varones), y el consumo perjudicial lleva a consecuencias físicas de la persona (mayor a 40g diarios en las mujeres y 60g diarios en varones). Los eventos excesivos de consumo de alcohol (episódico o circunstancial) resultan dañinos para ciertos problemas de salud (60g diarios de alcohol en una sola ocasión, o 5 bebidas estándar).

El termino bebida estándar es utilizado para simplificar la medición del consumo de alcohol, a pesar de ser inexacto. La OMS ha propuesto como valores de bebida estándar a:

Tipo de bebida	Cantidad o Equivalente
Cerveza (al 5%)	300ml de cerveza al 5% (1 lata, o a vaso de 220ml al 7%, 1 botella personal, o ½ botella mediana).
Vino (al 12%)	140ml de vino al 12% (una copa, 1 vaso pequeño o ½ vaso).
Aguardientes (al 40%)	40ml de bebida espirituosas (aguardientes) al 40% (1 trago).

El consumo de bebidas alcohólicas se refiere a ingesta de bebidas con contenido de alcohol (por ejemplo: cerveza, vino, aguardientes). Es una de las primeras causas de dependencia prevenible, y está asociado a otros problemas sociales y de salud como violencia, accidentes automovilísticos y trastornos de salud mental como depresión. La edad de inicio de consumo de alcohol está asociada al desarrollo de trastornos por uso de alcohol. Por esto es muy importante desarrollar medidas para retrasar el inicio de consumo de alcohol en los adolescentes.

Para la investigación, el consumo de bebidas alcohólicas es haber tomado al menos un vaso, una copa o una unidad similar. No se considera cuando consumió uno o dos "sorbos" o "bocados".

2.4. MODELOS DE ELECCIÓN DISCRETA

Los modelos de elección discreta son modelos en los que la variable dependiente es de carácter cualitativo (no métrica). Los modelos de elección discreta están muy relacionados con el análisis discriminante. Actualmente existe una tendencia a utilizar en mayor medida los modelos de elección discreta debido a que requieren realizar menos supuestos, lo que permite, en general obtener unos resultados más robustos.

En estadística, la regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados (GLM por sus siglas en inglés) que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelan, como una función de variables explicativas, utilizando una función logística.

La regresión logística es usada extensamente en las ciencias médicas y sociales. Otros nombres para regresión logística usados en varias áreas de aplicación incluyen modelo logístico, modelo logit, y clasificador de máxima entropía. La Regresión Logística es una técnica estadística multivariante que nos permite estimar la relación existente entre una variable dependiente no métrica, en particular dicotómica y un conjunto de variables independientes métricas o no métricas.

2.4.1. Introducción de Modelo Logístico Binario

Muchas situaciones dentro del análisis de datos involucran la predicción del valor de resultado de una variable dependiente categórica. Estas incluyen aplicaciones en la medicina para predecir el estado de salud de un paciente, en estudios de mercados la predicción acerca de la aceptación de un producto, en el ambiente pedagógico para predecir el rendimiento académico de un estudiante. La regresión logística binaria es una técnica que puede resultar bastante útil en estas y otras situaciones.

La regresión logística está diseñada para emplear una mezcla de variables predictoras categóricas y continuas para predecir una variable categórica de resultado o dependiente. Es muy común verlo como una alternativa al análisis discriminante. Varios de los conceptos que se discuten dentro de este análisis de regresión, son parte también de esa técnica multivariante, lo cual significa un complemento de estudio importante en la aplicación de la ciencia estadística.

A diferencia del análisis discriminante la regresión logística tiene muy pocos y menos estrictos, supuestos, además aun cuando los supuestos del análisis discriminante se han cumplido, la regresión logística se desempeña casi igual de bien. En particular este estudio hace referencia clara cuando se tiene una variable dependiente con dos categorías, y es allí cuando toma el nombre de regresión logística binaria.

Los modelos de regresión logística son modelos estadísticos en los que se desea conocer la relación entre: Una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos valores (regresión logística multinomial), en relación con una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas, siendo la ecuación inicial del modelo de tipo exponencial, si

bien su transformación logarítmica (logit) permite su uso como una función lineal.

Como se ve, las covariables pueden ser cuantitativas o cualitativas. Las covariables cualitativas deben ser dicotómicas, tomando valores 0 para su ausencia y 1 para su presencia (esta codificación es importante, ya que cualquier otra codificación provocaría modificaciones en la interpretación del modelo). Pero si la covariable cualitativa tuviera más de dos categorías, para su inclusión en el modelo debería realizarse una transformación de la misma en varias covariables cualitativas dicotómicas ficticias o de diseño (las llamadas variables dummy), de forma que una de las categorías se tomaría como categoría de referencia. Con ello cada categoría entraría en el modelo de forma individual. En general, si la covariable cualitativa posee n categorías, habrá que realizar n -1 covariables ficticias.

Pero, del conjunto de variables que pueda tener un estudio, ¿qué variables deben introducirse en el modelo? El modelo debe ser aquél más reducido que explique los datos (principio de parsimonia), y que además sea congruente e interpretable. Hay que tener en cuenta que un mayor número de variables en el modelo implicará mayores errores estándar. Deben incluirse todas aquellas variables que se consideren importantes para el modelo, con independencia de que si un análisis univariado previo demostró o no su significación estadística.

Por otro lado, no debería dejarse de incluir toda variable que en un análisis univariado previo demostrara una relación "suficiente" con la variable dependiente. Como se ve, no se habla de significación estadística ($p < 0,05$), que sería un criterio excesivamente restrictivo, sino de un cierto grado de relación (por ejemplo $p < 0,25$). La laxitud de esta recomendación se debe a que un criterio tan restrictivo como una $p < 0,05$; puede conducir a dejar de incluir en el modelo covariables con una débil asociación a la variable

dependiente en solitario pero que podrían demostrar ser fuertes predictores de la misma al tomarlas en conjunto con el resto de covariables.

El Análisis de Regresión Logística tiene la misma estrategia que el Análisis de Regresión Lineal Múltiple, el cual se diferencia esencialmente del Análisis de Regresión Logística por que la variable dependiente es métrica; en la práctica el uso de ambas técnicas tienen mucha semejanza, aunque sus enfoques matemáticos son diferentes.

La variable dependiente o respuesta no es continua, sino discreta (generalmente toma valores 1,0). Las variables explicativas pueden ser cuantitativas o cualitativas; y la ecuación del modelo no es una función lineal de partida, sino exponencial; si bien, por sencilla transformación logarítmica, puede finalmente presentarse como una función lineal. Así pues el modelo será útil en frecuentes situaciones prácticas de investigación en que la respuesta puede tomar únicamente dos valores: 1, presencia (con probabilidad p); y 0, ausencia (con probabilidad 1-p).

El modelo será de utilidad puesto que, muchas veces, el perfil de variables puede estar formado por caracteres cuantitativos y cualitativos; y se pretende hacer participar a todos ellos en una única ecuación conjunta.

El modelo puede acercarse más a la realidad ya que muchos fenómenos, como los del campo epidemiológico, se asemejan más a una curva que a una recta. Además la curva exponencial elegida como mejor ajuste, puede ser transformada logarítmicamente en una ecuación lineal de todas las variables, siendo así que el aparato matemático estudiado para la regresión lineal múltiple será aplicable; aunque el investigador tenga, al final, que deshacer la transformación para interpretar sus conclusiones.

Si para el Modelo de Regresión Logística una variable regresora de tipo categórica tiene c niveles habrá que generar c-1 variables ficticias (dummy) a fin que todas las posibilidades de la variable queden bien representadas en el modelo logístico. Cuando todas las variables regresoras son categóricas entonces se usa el modelo Log lineal, ver Mc Cullagh (1983).

Medina (2003), en su investigación sobre modelos de elección discreta dice; la utilidad de los modelos de elección discreta frente a la econometría tradicional radica en que los primeros permiten la modelización de variables cualitativas, a través del uso de técnicas propias de las variables discretas. Se dice que una variable es discreta cuando está formada por un número finito de alternativas que miden cualidades. Esta característica exige la codificación como paso previo a la modelización, proceso por el cual las alternativas de las variables se transforman en códigos o valores cuánticos, susceptibles de ser modelizados, utilizando técnicas econométricas.

Entre los modelos de elección discreta más conocidos se tienen:

Modelo	Descripción
Modelos dicotómicos o binomiales	Cuando la variable dependiente toma solamente dos modalidades o categorías. Las dos modalidades deben ser mutuamente excluyentes.
Modelos Multinomiales	Cuando la variable dependiente toma más de dos modalidades que son diferentes, exhaustivas y mutuamente excluyentes.
Modelos Ordenados	Cuando la variable dependiente toma más de dos modalidades, que son también, diferentes, exhaustivas y mutuamente excluyentes pero que, a diferencia de los multinomiales puede establecerse un orden, es decir, es una variable ordinal.
Modelos Loglineales	Estos modelos se utilizan para analizar tablas de contingencia de dos o más dimensiones.

2.4.2. Modelo Lineal de Probabilidad (MLP)

2.4.2.1. Especificaciones e interpretación del MLP

La primera tentativa teórica desarrollada para estudiar modelos con variables dicotómicas se planteó como una mera extensión del Modelo Lineal General que viene expresado por:

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e_i \qquad (1)$$

Dónde:

Y_i: puede tomar dos valores: 1 si ocurre una alternativa y 0 en caso contrario, de esta manera la variable dependiente no es continúa.

Xi: Variables independientes o explicativas

e_i : Variable aleatoria que se distribuye normal $N(0, \sigma^2)$

La distribución de la muestra en este tipo de modelos se caracteriza por configurar una nube de puntos de tal manera que las observaciones muestrales se dividen en dos subgrupos. Uno de ellos está formado por las observaciones en las que ocurrió el acontecimiento objeto de estudio (Y_i=1), y el otro, por los puntos muestrales en los que no ocurrió (Y_i=0).

El Modelo Lineal de Probabilidad (1), se puede interpretar en términos probabilísticos, en el sentido de que un valor concreto de la recta de regresión mide la probabilidad de que ocurra el acontecimiento objeto de estudio. Es decir, $\hat{Y}_i$ se puede considerar como la estimación de la probabilidad de que ocurra el acontecimiento objeto de estudio (Y_i=1) bajo el siguiente criterio: valores próximos a cero se corresponden con una baja probabilidad de ocurrencia del acontecimiento analizado (menor cuanto más próximos a cero);

mientras que a valores próximos a uno se les asigna una probabilidad elevada de ocurrencia (mayor cuanto más próximos a uno).

La interpretación de los coeficientes estimados en los Modelos Lineales de Probabilidad (MLP) es la misma que la del Modelo Lineal General, recogiendo el valor del parámetro el efecto de una variación unitaria en cada una de las variables explicativas sobre la probabilidad de ocurrencia del acontecimiento objeto de estudio. Así, si se produce un incremento de una unidad en la variable explicativa X_i ese aumento provocaría una variación igual a β_1 en la probabilidad $f_i(1)$.

2.4.2.2. Limitaciones de la estimación por MCO

La estimación del modelo anterior por Mínimos Cuadrados Ordinarios plantea una serie de limitaciones que se detallan a continuación:

1. El valor estimado puede estar fuera de rango [0 – 1], La estimación del Modelo Lineal de Probabilidad a través de MCO no garantiza que los valores estimados de Y_i estén entre 0 y 1, lo cual carece de lógica al interpretarse el valor estimado como una probabilidad. Este problema se soluciona truncando el rango de variación del valor estimado, dando lugar al modelo conocido con el nombre de Modelo Probabilístico Lineal Truncado, y que, para una única variable explicativa, se expresa de la forma:

$$Y_i \begin{cases} 1 & \alpha + \beta_{ki} X_{ki} \geq 1 \\ \alpha + \beta_{ki} X_{ki} & 0 < \alpha + \beta_{ki} X_{ki} < 1 \\ 0 & \alpha + \beta_{ki} X_{ki} \leq 0 \end{cases}$$

Sin embargo, si se restringen los valores de Y_i a 0 y 1, los valores del término independiente y la pendiente varían según los valores de Xi, de tal forma que:

Para $X_i \leq \frac{-\alpha}{\beta}$ término independiente y pendiente iguales a 0.

Para $\frac{-\alpha}{\beta} \leq X_i \leq \frac{(1-\alpha)}{\beta}$ término independiente igual a α y pendiente igual a β.

Para $X_i \geq \frac{(1-\alpha)}{\beta}$ término independiente igual a 1 y pendiente igual a 0.

Esto hará que si se incluyen en la estimación puntos en los que $X_i \leq \frac{-\alpha}{\beta}$ ó $X_i \geq \frac{(1-\alpha)}{\beta}$ los estimadores serán sesgados e inconsistentes.

2. La perturbación aleatoria puede no seguir una distribución Normal, dados los valores que toma la perturbación aleatoria no se puede asegurar que ésta se distribuya como una normal, al tratarse de una distribución binaria o dicotómica. Si bien el incumplimiento de la hipótesis de normalidad no invalida la estimación por MCO, sin embargo, la ausencia de normalidad imposibilita el uso de los estadísticos habituales utilizados para realizar el contraste de hipótesis tales como la t – Student, la F – Snedecor, etc. al basarse dichos contrastes en la hipótesis de normalidad de la perturbación aleatoria.

3. Problemas de Heterocedasticidad, aun en el caso de que se cumpliesen las hipótesis de media y correlación nula en la perturbación aleatoria $(E(\varepsilon_i) = 0 \; y \; E(\varepsilon_i \varepsilon_j) = 0 \;\; \forall \, i \neq j)$ no se cumple la hipótesis de varianza constante, es decir, la perturbación aleatoria no es homocedástica. Para comprobarlo se calcula la varianza de la perturbación aleatoria a través de su definición.

$$\begin{aligned} \mathrm{Var}(\varepsilon_i) = \mathrm{E}(\varepsilon_i - \mathrm{E}(\varepsilon_i))^2 &= \mathrm{E}(\varepsilon_i)^2 \\ &= (1 - \alpha - \beta_k X_{ki})^2 f_i(1) + (-\alpha - \beta X_{ki})^2 (1 - f_i(1)) \\ &= (1 - f_i(1))^2 f_i(1) + (f_i(1)^2 (1 - f_i(1)) \\ &= (1 - f_i(1)) f_i(1) (1 - f_i(1) + f_i(1)) \\ &= (1 - f_i(1)) f_i(1) \end{aligned}$$

La varianza de la perturbación aleatoria es una función de la probabilidad $f_i(1)$, la cuales a su vez, función de cada una de las observaciones de las variables explicativas X_i. La perturbación aleatoria es, por tanto, heterocedástica y la estimación del modelo mediante el método de MCO obtiene unos estimadores de los coeficientes de regresión con varianza no mínima, es decir, no eficientes.

Este problema podría solucionarse estimando el modelo a través de Mínimos Cuadrados Generalizados (MCG). A este tipo de modelos se les denomina Modelos Lineales Probabilísticos Ponderados. La estimación a través de MCG requiere la realización de los siguientes pasos:

- Se estima el modelo (1) mediante MCO sin tener en cuenta el problema de Heterocedasticidad, obteniéndose el valor estimado $\hat{y}_i$.

- El valor $\hat{y}_i$ se utiliza para calcular la varianza de la perturbación aleatoria, a través de la fórmula anteriormente obtenida:

$$Var(\varepsilon_i) = \left(1 - f_i(1)\right) f_i(1) = \hat{Y}_i\left(1 - \hat{Y}_i\right) = \sigma_i^2$$

- Si los valores estimados de $\hat{Y}_i$ son mayores que la unidad o menores que cero, deben sustituirse por la unidad (en el primer caso) o por cero (en el segundo). En ambos casos el valor resultante del cálculo de la varianza de e_i será cero, lo que generaría problemas al utilizar la $Var(\varepsilon_i)$ como ponderador. Ante esta situación se puede optar por eliminar las observaciones que generan estos valores, incurriendo en pérdida de información. Es por ello que la opción preferida es sustituir los valores mayores o iguales a la unidad por 0,999, y los valores menores o iguales a cero por 0,001.

- Se pondera el modelo (1) dividiendo ambos miembros de la ecuación por la desviación típica estimada $\sqrt{\sigma_i^2} = \sqrt{\hat{Y}_i(1-\hat{Y}_i)}$ con el fin de transformar el modelo en homocedástico.

$$\frac{Y_i}{\sqrt{\sigma_i^2}} = \beta_1 \frac{1}{\sqrt{\sigma_i^2}} + \beta_2 \frac{X_{1i}}{\sqrt{\sigma_i^2}} + \cdots + \beta_i \frac{X_{ki}}{\sqrt{\sigma_i^2}} + \varepsilon_i \frac{1}{\sqrt{\sigma_i^2}}$$

La estimación por MCO del modelo transformado es equivalente a aplicar MCG en el modelo (1) y en ambos casos se obtienen estimaciones eficientes de los coeficientes de regresión.

Sin embargo, uno de los problemas que presenta la estimación por MCG es la pérdida del término independiente en el modelo. La omisión del término independiente puede provocar que la suma de los residuos sea distinta de cero lo que puede tener consecuencias sobre el coeficiente de determinación (puede ser negativo), la función de verosimilitud estimada a partir de los residuos y los estadísticos que se obtienen a partir de ella.

4. El Coeficiente de Determinación R^2 esta subestimado, la suma de los cuadrados de los residuos ($\sum e_i^2$) es más grande de lo habitual debido a la forma específica en que se distribuye la nube de puntos de una variable dicotómica. Dado que el cálculo del coeficiente de determinación R^2 se ve afectado por $\sum e_i^2$, el R^2 calculado en la estimación por MCO es más pequeño de lo que realmente debería ser.

2.4.3. Modelos de probabilidad No lineal

La estimación e interpretación de los modelos probabilísticos lineales plantea una serie de problemas que han llevado a la búsqueda de otros modelos alternativos que permitan estimaciones más fiables de las variables

dicotómicas. Para evitar que la variable endógena estimada pueda encontrarse fuera del rango [0; 1], las alternativas disponibles son utilizar modelos de probabilidad no lineales, donde la función de especificación utilizada garantice un resultado en la estimación comprendido en el rango [0; 1]. Las funciones de distribución cumplen este requisito, ya que son funciones continuas que toman valores comprendidos entre 0 y 1.

2.4.3.1. Especificación de los modelos de elección discreta Logit y Probit

Dado que el uso de una función de distribución garantiza que el resultado de la estimación esté acotado entre 0 y 1, en principio las posibles alternativas son varias, siendo las más habituales la función de distribución logística, que ha dado lugar al modelo Logit, y la función de distribución de la normal tipificada, que ha dado lugar al modelo Probit. Tanto los modelos Logit como los Probit se relacionan, por tanto, la variable endógena Yi con las variables explicativas X_i a través de una función de distribución.

En el caso del modelo Logit, la función utilizada es la logística, por lo que la especificación de este tipo de modelos queda como sigue:

$$Y_i = \frac{1}{1 + e^{-\alpha - \beta_{ki} X_{ki}}} + \varepsilon_i = \frac{e^{\alpha + \beta_{ki} X_{ki}}}{1 + e^{\alpha + \beta_{ki} X_{ki}}} + \varepsilon_i \qquad (2)$$

En el caso del modelo Probit la función de distribución utilizada es la de la normal tipificada, con lo que el modelo queda especificado a través de la siguiente expresión

$$Y_i = \int_{-\infty}^{\alpha + \beta X_i} \frac{1}{(2\pi)^{1/2}} e^{-\frac{z^2}{2}} dz + \varepsilon_i \qquad (3)$$

Donde la variable z es una variable "muda" de integración con media cero y varianza uno.

Dada la similitud existente entre las curvas de la normal tipificada y de la logística, los resultados estimados por ambos modelos no difieren mucho entre sí, siendo las diferencias operativas, debidas a la complejidad que presenta el cálculo de la función de distribución normal frente a la logística, ya que la primera solo puede calcularse en forma de integral. La menor complejidad de manejo que caracteriza al modelo Logit es lo que ha potenciado su aplicación en la mayoría de los estudios empíricos.

Al igual que en el Modelo Lineal de Probabilidad, el Modelo Logit se puede interpretar en términos probabilísticos, es decir, sirve para medir la probabilidad de que ocurra el acontecimiento objeto de estudio (Y_i=1). En cuanto a la interpretación de los parámetros estimados en un modelo Logit, el signo de los mismos indica la dirección en que se mueve la probabilidad cuando aumenta la variable explicativa correspondiente, sin embargo, la cuantía del parámetro no coincide con la magnitud de la variación en la probabilidad (como si ocurría en el MLP). En el caso de los modelos Logit, al suponer una relación no lineal entre las variables explicativas y la probabilidad de ocurrencia del acontecimiento, cuando aumenta en una unidad la variable explicativa los incrementos en la probabilidad no son siempre iguales ya que dependen del nivel original de la misma.

2.4.4. La Ecuación Logística

La regresión logística es un instrumento estadístico de análisis bivariado o multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia se ha puntuado con los valores cero y uno, respectivamente) y un conjunto de m variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas. En este último caso, se requiere que sean transformadas en variables ficticias o simuladas ("dummy"). El propósito del análisis es:

- Predecir la probabilidad de que a alguien le ocurra cierto evento: por ejemplo, "estar desempleado" =1 o "no estarlo" = 0; "ser pobre" = 1 o "no ser pobre" = 0; "graduarse como sociólogo" =1 o "no graduarse" = 0.

- Determinar qué variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión.

Esta asignación de probabilidad de ocurrencia del evento a un cierto sujeto, así como la determinación del peso que cada una de las variables independientes tienen en esta probabilidad, se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos.

La regresión logística binaria (a la que desde ahora solo se llamara regresión logística) es una regresión que se aplica con una variable independiente dicotómica, donde la variable dependiente no contiene valores de datos sin procesar, pero en cambio es la oportunidad de que en un evento de interés ocurra. En términos generales la ecuación de la regresión logística es:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta_k X_{ki} + \varepsilon_i$$

Donde los términos de la derecha son los términos estándar para las variables independientes y la intercepción es una ecuación de regresión, sin embargo, del lado izquierdo está el logaritmo natural de la oportunidad y la cantidad ln (Odds) llamada logit. En principio puede variar de menos a más infinito, por lo que se elimina el problema de la predicción fuera de los límites de la variable dependiente. La oportunidad está relacionada con la probabilidad por:

$$Odds = \frac{P_i}{1-P_i}$$

Vea que hay una relación lineal con las variables independientes en la regresión logística, pero es lineal en el logaritmo natural de la oportunidad y no en las probabilidades originales dado que estamos interesados en la probabilidad de un evento, por ejemplo el código más alto de una variable dicotómica o categoría de interés, la ecuación logística se puede transformar en otra ecuación en la probabilidad, entonces tiene esta forma:

$$\boldsymbol{Prob(evento\ Y_i = 1) = \frac{1}{1 + e^{-(\alpha + \beta_{ki} X_{ki})}}}$$

Dónde: Prob (y =1 | X) es la probabilidad de que y tome el valor 1 (presencia de la característica estudiada), en presencia de las covariables X:

$\boldsymbol{X_{ki}}$: es un conjunto de k covariables que forman parte del modelo.

$\boldsymbol{\alpha}$: es la constante del modelo o término independiente.

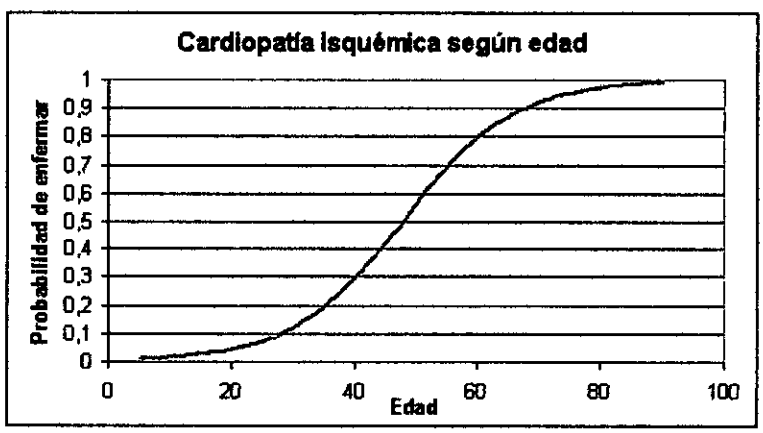
$\boldsymbol{\beta_{ki}}$: Los coeficientes de las covariables.

Esta ecuación no se puede calcular con el método de mininos cuadrados, en lugar de esto, los parámetros del modelo se estiman con la técnica de máxima verosimilitud. Derivamos los coeficientes que hacen que nuestros valores observados sean los más probables para el grupo de las variables independientes, esto se hace a través de iteraciones internas dentro de los programas estadísticos. La probabilidad que se obtiene es el punto de partida para la clasificación de un elemento cualquiera de la muestra a uno de los grupos definidos dentro de la variable de respuesta.

Los modelos de regresión logística binaria resultan los de mayor interés ya que la mayor parte de las circunstancias analizadas en el campo de la investigación responden a este modelo (presencia o no de enfermedad, éxito o fracaso, etc.). Como se ha visto, la variable dependiente será una variable dicotómica que se codificará como 0 ó 1 (respectivamente, "ausencia" y "presencia"). Este aspecto de la codificación de las variables no es banal (influye en la forma en que se realizan los cálculos matemáticos), y habrá que tenerlo muy en cuenta si se emplean paquetes estadísticos que no recodifican

automáticamente las variables cuando éstas se encuentran codificadas de forma diferente (por ejemplo, el uso frecuente de 1 para la presencia y –1 ó 2 para la ausencia).

La ecuación de partida en los modelos de regresión logística es lo que se denomina distribución logística. En la siguiente imagen vemos un ejemplo de esta distribución: la probabilidad de padecer enfermedad coronaria en función de la edad. Como puede verse, la relación entre la variable dependiente (cualitativa dicotómica), y la covariable (edad, cuantitativa continua en este caso), no es definida por una recta (lo que correspondería un modelo lineal), sino que describe una forma sigmoidea (distribución logística).



2.4.5. Elementos del Análisis de Regresión Logística

Cuando se hace un análisis de regresión logística se tiene dos objetivos generales:

1. Determinar el efecto de un grupo de variables en la probabilidad, además del efecto de las variables tomadas individualmente sobre el resultado global del modelo. Esto significa la ejecución de cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente, lo que lleva implícito también clarificar la existencia de interacción y

confusión entre covariables respecto a la variable dependiente (es decir, conocer los Odds ratio para cada covariable).

2. Alcanzar la más alta precisión predictiva que sea posible con un determinado grupo de variables predictoras seleccionadas a partir de la significancia obtenida por cada una de ellas. También se asume como consecuencia la acción de clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables.

Estos dos objetivos no son incompatibles, pero uno como el otro tiende a enfocarse más en un análisis específico. Quienes están interesados en la teoría y en los efectos causales suelen estar más preocupados en el primer objetivo, mientras lo que están interesados en la predicción y la aplicación funcional del modelo sobre una base de datos en el contexto real se enfocaran más en el segundo objetivo. No cabe duda que la regresión logística es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en investigación, de ahí su amplia utilización. El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías.

Varios de los pasos de la regresión logística son similares a la regresión estándar. Primero debe seleccionarse un grupo razonable de variables predictoras y deben revisarse los datos muy bien, antes de buscar patrones poco usuales; también debe generarse un estudio exploratorio descriptivo para ver los problemas de valores perdidos, erróneos o vacíos de cada variable independiente. Después de estimar la ecuación y revisar el efecto de cada variable de forma individual, debe hacerse la verificación para ver si los datos cumplen los supuestos del modelo logístico y buscar los casos que tienen

influencia excesiva en los resultados. Si está interesado en predecir la pertenencia a una categoría de interés de cierto evento particular es muy importante que el modelo se valide, fragmentado una sub muestra de aproximadamente el 25% del total de data utilizada y probándolo luego el modelo en la muestra restante.

Desde luego que hay diferencias con la regresión múltiple lineal a partir de las tablas de resultados, en la regresión estándar hay más de una forma de medir el ajuste del modelo y más de una forma de medir la cantidad de varianza explicada, es por ello que más adelante se muestran valores alternos a estos indicadores.

Como podemos afirmar en el análisis discriminante, la bondad de ajuste o significancia de un modelo no necesariamente equivale a una elevada precisión predictiva. Conforme aumenta el tamaño de la muestra un grupo de variables independientes pueden ser estadísticamente significativas, pero aun así no permitir un porcentaje elevado de predicciones correctas.

La clasificación de casos es un asunto simple en la regresión logit, se predice que un caso está en el valor más bajo de la variable dependiente si su probabilidad predicha es menor que 0.5, de otra forma se predice que está en la categoría superior.

2.4.6. Supuestos de la Regresión Logística

La regresión logit, tiene menos supuestos que la regresión estándar, la logística necesita que:

1. Las variables independientes sean intervalares, de razón o dicotómicas.

2. Que se incluya a todas las predictoras relevantes, que no incluya variables irrelevantes y que tengan una relación lineal entre sí.

3. El valor esperado del error es 0.

4. Que no haya autocorrelación.

5. Que no haya correlación entre el error y las variables independientes.

6. Que no haya multicolinealidad entre las variables independientes.

Multicolinealidad: Si bien existen pruebas que permiten comprobar la existencia de colinealidad entre covariables, cabe reseñar aquí que los modelos con multicolinealidad entre las covariables introducidas llamarán la atención por la presencia de grandes errores estándar, y frecuentemente, estimaciones de coeficientes anormalmente elevadas. Sin embargo la multicolinealidad no afecta al sentido de las estimaciones (la multicolinealidad no hará que aparezca significación donde no la hay, y viceversa).

Se dice que existe multicolinealidad cuando dos o más de las covariables del modelo mantienen una relación lineal. Cuando la colinealidad es perfecta, es decir, cuando una covariable puede determinarse según una ecuación lineal de una o más de las restantes covariables, es posible estimar un único coeficiente de todas las covariables implicadas. En estos casos debe eliminarse la covariable que actúa como dependiente.

Normalmente lo que se hallará será una multicolinealidad moderada, es decir, una mínima correlación entre covariables. Si esta correlación fuera de mayor importancia, su efecto sería, como ya se vio anteriormente, el incremento exagerado de los errores estándar, y en ocasiones, del valor estimado para los coeficientes de regresión, lo que hace las estimaciones poco creíbles. Un primer paso para analizar este aspecto puede ser examinar la matriz de coeficientes de correlación entre las covariables. Coeficientes de correlación muy elevados llevarán a investigar con mayor profundidad. Sin embargo, este

método, bueno para detectar colinealidad entre dos covariables, puede conducir a no poder detectar multicolinealidad entre más de dos de ellas.

Existen otros procedimientos analíticos para detectar multicolinealidad. Puede desentenderse por el momento de la variable dependiente y realizar sendos modelos en los que una de las covariables actuará como variable dependiente y las restantes covariables como variables independientes de aquella.

Los dos supuestos siguientes se hacen en la regresión estándar pero no en la logística.

1. Normalidad de los errores, se asume que los errores siguen distribución binomial y se solo se aproxima a una normal cuando la muestra es muy grande.

2. Homogeneidad de varianza, como discutimos antes esta condición no se puede mantener por definición.

A diferencia del análisis discriminante el empleo de un gran número de variables dummy como predictoras no viola ningún supuesto de regresión logit, por lo que puede preferirse a esta en este tipo de situaciones. Se debe mencionar que si las circunstancias fueran las mismas que en la regresión estándar, la regresión logística necesita de tamaños de muestra grande para su aplicación, los teóricos y conocedores dicen que debe utilizarse tamaños entre 10 y 30 veces el número de variables independientes para lograr una buena inferencia.

La diferencia básica entre los modelos del Análisis de Regresión Lineal Múltiple y de la Regresión Logística es naturaleza de la relación entre la variable respuesta y las variables regresoras. Para el Análisis de Regresión Lineal Múltiple, consideremos Y una variable respuesta cuantitativa y X_1, X_2,..... , X_k variables regresoras o llamadas también explicativas; y se desea

describir la relación que hay entre la variable respuesta y las variables explicativas, si entre la variable respuesta y las regresoras hay una relación lineal se espera que dicha función pueda ser expresada en términos cuantitativos.

Los Estadísticos Odds Ratio:

El Estadístico Odds: mide el cociente de probabilidades para una observación i de elegir la opción 1 frente a la opción 0, es decir:

$$Odds = \frac{P_i}{1 - P_i}$$

Este indicador en términos de regresión logística es igual a:

$$Odds = \frac{P_i}{(1 - P_i)} = e^{\alpha + \beta_i X_i}$$

El Estadístico Odds Ratio: Se define como el cociente de Odds de esta manera se tiene:

$$Odds\ Ratio = \frac{\frac{P_i}{(1 - P_i)}}{\frac{P_m}{1 - P_m}}$$

Si lo que se quiere es comparar la utilidad que la opción elegida proporciona al individuo (observación) i, con respecto a la utilidad percibida por el individuo (observación) m, entonces se define como Odds Ratio.

De este modo en términos de regresión logística es:

$$Odds\ Ratio = \frac{\frac{P_{i+1}}{(1 - P_{i+1})}}{\frac{P_i}{(1 - P_i)}} = \frac{e^{\alpha + \beta_i (X_i + 1)}}{e^{\alpha + \beta_i X_i}} = e^{\beta_i (X_i + 1 - X_i)} = e^{\beta_i}$$

2.4.7. Estimación de los Parámetros en los Modelos Logit

Antes de abordar el método de estimación en los modelos Logit, es preciso distinguir la existencia de dos casos diferenciados que implican la utilización de métodos de estimación distintos: los modelos Logit con observaciones repetidas y con observaciones no repetidas. Para la estimación de los coeficientes del modelo y de sus errores estándar se recurre al cálculo de estimaciones de máxima verosimilitud, es decir, estimaciones que hagan máxima la probabilidad de obtener los valores de la variable dependiente Y proporcionada por los datos de nuestra muestra. Estas estimaciones no son de cálculo directo, como ocurre en el caso de las estimaciones de los coeficientes de regresión de la regresión lineal múltiple por el método de los mínimos cuadrados. Para el cálculo de estimaciones máximo–verosímiles se recurre a métodos iterativos, como el método de Newton–Raphson.

Dado que el cálculo es complejo, normalmente hay que recurrir al uso de rutinas de programación o paquetes estadísticos. De estos métodos surgen no sólo las estimaciones de los coeficientes de regresión, sino también de sus errores estándar y de las covarianzas entre las covariables del modelo.

Para el caso sencillo de una única variable explicativa, nos encontramos en una situación con observaciones repetidas cuando la variable X es discreta y presenta un número reducido de alternativas o intervalos (F), de manera que para cada alternativa de la variable X tendremos n_i observaciones de Y, pudiéndose calcular las proporciones o probabilidades muestrales. En este caso la matriz de n datos muestrales queda reducida a F observaciones siendo los valores que tome la variable dependiente (π_i) las proporciones muestrales calculadas a través de la expresión:

$$\pi_i = \sum_{i=1}^{F} \frac{Y_i}{n_i}$$

La generalización del modelo a k variables explicativas implica la existencia de observaciones repetidas de Y para cada combinación de las k variables explicativas, pudiéndose calcular las proporciones o probabilidades muestrales de la misma forma que en el caso anterior. En este caso, si bien los valores de la variable dependiente están acotados en el rango 0-1, son valores continuos, por lo que el método utilizado para la estimación de los parámetros del modelo es el que habitualmente se trabajó con variables continuas. Por lo tanto, ante la presencia de observaciones repetidas, se podría aplicar el método de Mínimos Cuadrados Ordinarios. Sin embargo, la existencia de Heterocedasticidad en el modelo obliga a estimar por Mínimos Cuadrados Generalizados, para garantizar el cumplimiento de las propiedades de los parámetros estimados, utilizándose la inversa de la varianza de los errores como ponderación del modelo.

Sin embargo, lo más habitual es no poder calcular las probabilidades muestrales, bien porque las variables independientes incluidas en el modelo son continuas, o bien porque aun siendo éstas discretas, la combinación de las mismas impide la obtención de observaciones repetidas de la variable dependiente para cada uno de los intervalos F. En esta situación, la matriz de datos muestrales estará formada por n observaciones pudiendo ser el valor de la variable dependiente para cada una de ellas 1 ó 0. La naturaleza dicotómica de la variable dependiente en este tipo de modelos impide la utilización de los métodos tradicionales en la estimación de los parámetros, al no poderse calcular la inversa de la varianza utilizada como ponderación del modelo. Para la estimación de los parámetros se utiliza el método de Máxima Verosimilitud.

A continuación se describen ambos métodos de estimación (máxima verosimilitud y mínimos cuadrados generalizados) comenzando por el caso más habitual de ausencia de observaciones repetidas.

2.4.7.1. Estimación con observaciones no repetidas: Método de Máxima Verosimilitud

Dada una variable aleatoria, caracterizada por unos parámetros, y dada una muestra poblacional, se consideran estimadores Máximo-Verosímiles de los parámetros de una población determinada, aquellos valores de los parámetros que generarían con mayor probabilidad la muestra observada. Es decir, los estimadores Máximo-Verosímiles son aquellos valores para los cuales la función de densidad conjunta (o función de verosimilitud) alcanza un máximo.

Suponiendo que las observaciones son independientes, la función de densidad conjunta de la variable dicotómica Y_i queda como:

$$Prob(Y_1 Y_2 \dots Y_i \dots Y_n) = \pi_{i=1}^{n} P_i^{Y_i} (1 - P_i)^{1-Y_i}$$

Donde P_i recoge la probabilidad de que $Y_i = 1$. Por simplicidad se trabaja con la función de densidad conjunta en logaritmos, cuya expresión es:

$$\pounds = \ln L = \sum_{i=1}^{i} Y_i \ln P_i + \sum_{i=1+i}^{n-i} (1 - Y_i) \ln(1 - P_i)$$
$$= \sum Y_i \ln P_i + \sum (1 - Y_i) \ln(1 - P_i)$$

El método de estimación de máxima verosimilitud elige el estimador del parámetro que maximiza la función de verosimilitud ($\pounds = ln\, L$), por lo que el procedimiento a seguir será calcular las derivadas de primer orden de esta función con respecto a los parámetros que queremos estimar, igualarlas a 0 y resolver el sistema de ecuaciones resultante. Las derivadas de primer orden de la función de verosimilitud respecto a los parámetros α y β, tras pequeñas manipulaciones, quedan como siguen:

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_{i=1}^{n} (Y_i - P_i) = \sum (Y_i - \frac{e^{\hat{\alpha}+\hat{\beta}X_i}}{1+e^{\hat{\alpha}+\hat{\beta}X_i}}) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_{i=1}^{n} (Y_i - P_i)X_i = \sum (Y_i - \frac{e^{\hat{\alpha}+\hat{\beta}X_i}}{1+e^{\hat{\alpha}+\hat{\beta}X_i}})X_i = 0$$

Sustituyendo P_i por su valor queda:

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_{i=1}^{n} e_i = \sum (Y_i - \frac{e^{\hat{\alpha}+\hat{\beta}X_i}}{1+e^{\hat{\alpha}+\hat{\beta}X_i}}) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_{i=1}^{n} X_i e_i = \sum (Y_i - \frac{e^{\hat{\alpha}+\hat{\beta}X_i}}{1+e^{\hat{\alpha}+\hat{\beta}X_i}})X_i = 0$$

Se trata de un sistema de ecuaciones no lineales por lo que es necesario aplicar un método iterativo o algoritmo de optimización que permita la convergencia en los estimadores.

2.4.7.2. Estimación con observaciones repetidas: Método Mínimos Cuadrados Generalizados

La estimación del modelo con datos agrupados podría realizarse mediante el procedimiento habitual utilizado para estimar regresiones lineales, ya que la variable a modelizar ya no es dicotómica (es continua aunque acotada en el rango 0-1). Para ello es necesario linealizar el modelo, lo cual es fácil de realizar a través de la transformación ya comentada anteriormente, y por la cual:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta_k X_{ki} + \varepsilon_i$$

Donde ε_i es el valor de la perturbación aleatoria incluida en la especificación de todo modelo de regresión lineal y que cumple las hipótesis de perturbación esférica y ausencia de autocorrelacion. El modelo así transformado puede estimarse por el procedimiento

habitual de Mínimos Cuadrados Ordinarios (MCO). Sin embargo, y dado que el valor de P_i es desconocido y debe sustituirse por su estimación muestral $\hat{P}_i$, el modelo a estimar quedaría como:

$$\ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = \alpha + \beta_k X_{ki} + \varepsilon_i + \varepsilon_i'$$

Donde ε_i' recoge el error cometido al utilizar la estimación muestral de la probabilidad $\hat{P}_i$, en vez de su valor desconocido P_i. Al sustituir P_i por su estimación muestral $\hat{P}_i$, los errores, supuestos independientes, cumplen la condición asintótica de normalidad exigida para realizar contrastaciones y construcción de intervalos de confianza, pero dejan de cumplir la condición de homocedasticidad ya que su varianza no es constante.

La presencia de heterocedasticidad impide la estimación a través de Mínimos Cuadrados Ordinarios, siendo necesario aplicar el método de Mínimos Cuadrados Generalizados, que sin exigir la condición de homocedasticidad de los errores, permite estimar estimadores. Este procedimiento transforma el modelo a estimar en otro, donde todas las variables quedan ponderadas por los inversos de las varianzas de los errores, y dado que se desconocen dichos valores verdaderos, éstos se sustituyen por su estimación muestral P_i, de donde:

$$s_i = \frac{1}{\hat{V}ar(\varepsilon_i')} = n_i\hat{P}_i(1-\hat{P}_i)$$

Quedando el modelo a estimar como:

$$s_i Ln\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = \alpha s_i + \beta_k X_{ki} s_i + \varepsilon_i$$

Análisis de los Indicadores de la Varianza Explicada

La regresión logística también proporciona dos valores que son análogos, al coeficiente de determinación R cuadrado, de la regresión estándar. Dada la relación para una variable dicotómica, la cantidad de varianza explicada del modelo se debe definir diferente. La R cuadrado de Cox y Snell y la R cuadrado de Nagelkerke, por lo general se prefiere la segunda porque puede llegar a alcanzar un valor máximo de 1. Por cualquiera de estos dos valores su definición es acerca de lo que las variables independientes pueden alcanzar a explicar de la varianza total del modelo.

Su identificación está relacionada con la prueba F de significancia global del modelo, y esta se origina después de analizar el ajuste a través de las pruebas ómnibus especificas del modelo logit y los resultados obtenidos en ambas pruebas influyen directamente sobre los estudios futuros de la predicción.

2.4.8. Contraste y Validación de Hipótesis

En el caso de trabajar con observaciones repetidas la contrastación y validación del modelo estimado sigue la misma metodología que la empleada en el análisis de regresión tradicional, por lo que remitimos a éste para profundizar en este tema. Mientras que si nos encontramos en el caso de no disponer de observaciones repetidas, la etapa de contrastación y validación del modelo estimado por máxima-verosimilitud se lleva a cabo aplicando los estadísticos específicos que se comentan a continuación.

2.4.8.1. Significatividad estadística de los parámetros estimados

La distribución del estimador del parámetro β es aproximadamente $N(\beta; \sigma_{\hat{\beta}})$. En tal situación, se puede construir un intervalo de

confianza del parámetro estimado, para testar si dicho valor es significativamente distinto de cero de forma individual. El contraste a realizar quedaría definido como:

$H_0: \beta = 0$ El parámetro es igual a cero

$H_1: \beta \neq 0$ El parámetro es distinto de cero

El intervalo de confianza proporciona un rango de posibles valores para el parámetro, por lo que si el valor estimado no pertenece a dicho intervalo, se deberá rechazar la hipótesis nula. El intervalo quedaría definido como:

$$\hat{\beta} - Z_{\frac{\alpha}{2}}\sigma_{\hat{\beta}} \leq \beta \leq \hat{\beta} + Z_{\alpha/2}\sigma_{\hat{\beta}}$$

donde α es la probabilidad de que el verdadero valor del parámetro β se halle fuera del intervalo, y z es el valor tabular de la distribución N(0;1) que deja a su derecha una probabilidad igual a $\alpha/2$.

A partir de la expresión anterior se puede fijar un rechazo de la hipótesis nula cuando:

$$\left|\frac{\hat{\beta}}{\sigma_{\hat{\beta}}}\right| \geq Z_{\alpha/2}$$

2.4.8.2. Medidas de Bondad de Ajuste del modelo

Se sabe que cualquier variable dependiente de otra u otras variables, toma valores según las variables de las que depende. Por otra parte, esa variable dependiente irá tomando valores siguiendo o describiendo una determinada distribución de frecuencias; es decir, toman los valores que tienen las variables independientes, si el experimento se repite múltiples veces, la variable dependiente tomará para un conjunto de variables independientes un determinado valor, y la probabilidad de ocurrencia de dicho valor vendrá dado por una distribución de frecuencias concreta: una distribución normal, una

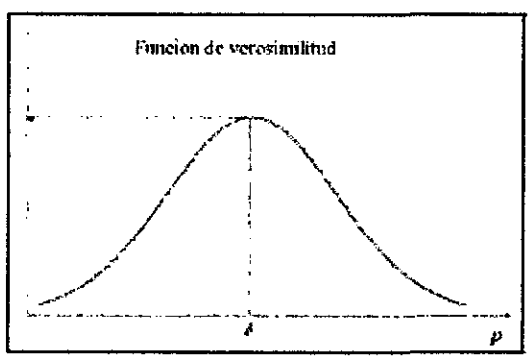
distribución binomial, una distribución Hipergeométrica, etc. En el caso de una variable dependiente dicotómica (como el caso que nos ocupa), la distribución de frecuencias que seguirá será la binomial, que depende de la tasa de éxitos (X sujetos de un total de N, que sería el elemento variable), para un determinado tamaño muestral N y probabilidad Pr (i) de ocurrencia del evento valorado por la variable dependiente (parámetros constantes). La función de densidad de esta distribución de frecuencias vendrá dada por la siguiente expresión:

$$\Pr(y) \approx f(x) = \binom{N}{x} p^x (1-p)^{N-x}$$

Si en la expresión anterior introducimos los datos concretos de nuestra muestra de N sujetos (es decir, se convierte el elemento variable X en parámetro), y se hace depender el resultado de la función de densidad del parámetro "probabilidad de ocurrencia" (p , que de esta forma se convierte en variable), se está generando su función de verosimilitud, $f(p \mid x)$ (función dependiente de p dado el valor muestral de x) o $L(p)$ (L de *likelihood*), que ofrece como resultados las probabilidades de la función de densidad ajustada a los datos:

$$f(p \mid x) = \binom{N}{x} p^x (1-p)^{N-x}$$

Se deduce que, para una *muestra* concreta, esa probabilidad será diferente según qué valores tome el parámetro "probabilidad de ocurrencia":



Se demuestra que la mejor estimación del parámetro es aquel valor que maximiza esta función de verosimilitud, ya que son estimadores consistentes (conforme crece el tamaño muestral, la estimación se aproxima al parámetro desconocido), suficientes (aprovechan la información de toda la muestra), asintóticamente normales y asintóticamente eficientes (con mínima varianza), si bien no siempre son insesgados (no siempre la media de las estimaciones para diferentes muestras tenderá hacia el parámetro desconocido).

El uso de la función de verosimilitud en la estimación, hace que la bondad del ajuste en los modelos de elección discreta sea un tema controvertido, ya que en estos modelos no existe una interpretación tan intuitiva como en el modelo de regresión clásico. A continuación se describen los contrastes más utilizados en la literatura para medir la bondad de ajuste en un modelo Logit y que concretaremos en: Estadístico de Wald, índice de cociente de verosimilitudes, el estadístico chi-cuadrado de Pearson, el porcentaje de aciertos estimados en el modelo, y la prueba de Hosmer-Lemeshow.

Estadístico de Wald:

Evalúa la significancia de los coeficientes, se define como el vector matriz de los coeficientes estimados del siguiente modo según las Hipótesis y donde se busca contrastar la proposición de que un coeficiente aislado es distinto de 0, y sigue una distribución normal de media 0 y varianza 1. Su valor para un coeficiente concreto viene dado por el cociente entre el valor del coeficiente y su correspondiente error estándar. La obtención de significación indica que dicho coeficiente es diferente de 0 y merece la pena su conservación en el modelo. En modelos con errores estándar grandes, el estadístico de Wald puede proporcionar falsas ausencias de significación (es decir, se incrementa el error tipo II). Tampoco es recomendable su uso si se están empleando variables de diseño.

$H_0: \beta_i = 0, \forall i$

$H_1: Al\ menos\ un\ \beta_i \neq 0$

Estadístico de Prueba:

$$W = \hat{\beta}' \left[\hat{I}(\hat{\beta})\right]^{-1} \hat{\beta} = \hat{\beta}' (X' V X) \hat{\beta} \sim X^2_{\alpha, k+1}$$

Donde X y V son las matrices:

$$X = \begin{bmatrix} 1 & x_{11} & \ldots\ldots & x_{1k} \\ 1 & x_{21} & \ldots\ldots & x_{2k} \\ \ldots & \ldots & \ldots\ldots & \ldots \\ 1 & x_{n1} & \ldots\ldots & x_{nk} \end{bmatrix}_{nx(k+1)}$$

$$V = \begin{bmatrix} \hat{P}(x_1)(1-\hat{P}(x_1)) & 0 & \ldots\ldots & 0 \\ 0 & \hat{P}(x_2)(1-\hat{P}(x_2)) & \ldots\ldots & 0 \\ \ldots & \ldots & \ldots\ldots & \ldots \\ 0 & 0 & \ldots\ldots & \hat{P}(x_n)(1-\hat{P}(x_n)) \end{bmatrix}_{nxn}$$

Decisión: si $W > X^2_{\alpha,k}$ rechazamos H_0 con un nivel de significancia fijado α, concluimos que la variable independiente influye en la probabilidad del suceso.

Índice de Cociente de Verosimilitudes:

La función de verosimilitud puede también utilizarse para obtener un estadístico, que tiene cierta semejanza con el coeficiente de determinación calculado en la estimación lineal, conocido "índice de cociente de verosimilitudes". Este estadístico compara el valor de la función de verosimilitud de dos modelos; uno corresponde al modelo estimado que incluye todas las variables independientes (modelo

completo) y el otro sería el del modelo cuya única variable independiente es la constante (modelo restringido). El estadístico, también conocido como R^2 de McFadden ya que fue propuesto por McFadden en 1974, se define como:

$$RV = ICV = 1 - \frac{\log L}{\log L(0)}$$

Donde L es el valor de la función de verosimilitud del modelo completo (el estimado con todas las variables independientes) y L(0) es el valor correspondiente del modelo restringido (el que incluye únicamente en la estimación el término constante).

Se trata de ir contrastando cada modelo que surge de eliminar de forma aislada cada una de las covariables frente al modelo completo. En este caso cada estadístico R.V. sigue una χ 2 con un grado de libertad (no se asume normalidad). La ausencia de significación implica que el modelo sin la covariable no empeora respecto al modelo completo (es decir, da igual su presencia o su ausencia), por lo que según la estrategia de obtención del modelo más reducido (principio de parsimonia), dicha covariable debe ser eliminada del modelo ya que no aporta nada al mismo. Esta prueba no asume ninguna distribución concreta, por lo que es la más recomendada para estudiar la significación de los coeficientes.

El ratio calculado tendrá valores comprendidos entre 0 y 1 de forma que:

- Valores próximos a 0 se obtendrán cuando L(0) sea muy parecido a L, situación en la que nos encontraremos cuando las variables incluidas en el modelo sean poco significativas, es decir, la estimación de los parámetros β no mejora el error que se comete si dichos parámetros se igualaran a 0. Por lo que en

este caso la capacidad explicativa del modelo será muy reducida.

- Cuanto mayor sea la capacidad explicativa del modelo, mayor será el valor de L sobre el valor de L(0), y más se aproximará el ratio de verosimilitud calculado al valor 1.

El estadístico χ^2 de Pearson:

Para medir la bondad del ajuste también se utilizan medidas del error que cuantifican la diferencia entre el valor observado y el estimado. En concreto, para contrastar la hipótesis nula de que:

$$H_0: Y_i = \hat{P}_i \; ; lo\ que\ equivale\ a\ H_0: Y_i - \hat{P}_i = e_i = 0$$

Se construye un estadístico que recoge los residuos estandarizados o de Pearson del modelo Logit, que se definen como la diferencia entre el valor observado de la variable respuesta y el estimado, dividido por la estimación de la desviación típica, ya que la esperanza es nula. A través del contraste de multiplicadores de Lagrange, se puede calcular el estadístico conocido con el nombre de χ^2 de Pearson, que se define como:

$$\chi^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{P}_i)^2}{\hat{P}_i(1-\hat{P}_i)}$$

Este estadístico es similar a la suma de cuadrados de los residuos del modelo de regresión convencional. El ajuste del modelo será mejor cuanto más cerca esté el valor del estadístico de cero. Para saber a partir de qué valor puede considerarse el ajuste como aceptable es necesario conocer la distribución del estadístico. Éste estadístico, bajo la hipótesis nula, se distribuye como una chi-cuadrado con (n-k) grados de libertad, por lo que su valor se compara con el valor teórico

de las tablas de la chi-cuadrado para contrastar la hipótesis nula. Si el valor calculado es superior al valor teórico se rechazara la hipótesis nula lo que equivale a decir que el error cometido es significativamente distinto de cero, es decir, se trataría de un mal ajuste.

Porcentaje de aciertos estimados del modelo:

Otra de las vías utilizadas para determinar la bondad de un modelo Logit es predecir con el modelo los valores de la variable dependiente Y_i de tal manera que $Y_i = 1$ si $\hat{P}_i > c$ ó $Y_i = 0$ si $\hat{P}_i > c$. Generalmente, el valor que se asigna a c para determinar si el valor de la predicción es igual a 1 o a 0 es de 0.5, puesto que parece lógico que la predicción sea 1 cuando el modelo dice que es más probable obtener un 1 que un 0.

Sin embargo, la elección de un umbral igual a 0.5 no siempre es la mejor alternativa. En el caso en que la muestra presente desequilibrios entre el número de unos y el de ceros la elección de un umbral igual a 0.5 podría conducir a no predecir ningún uno o ningún cero. Así, supuesta una muestra de 1000 observaciones donde 100 son 1 y el resto 0, si el modelo incluye término constante, la media de las probabilidades estimadas en la muestra será 0.1, se deduce que la media de las probabilidades estimadas por el modelo, ha de coincidir con la proporción de unos que haya en la muestra, por lo que será casi imposible que se obtenga un valor estimado superior a 0.5. Si el umbral seleccionado es de 0.5, con esta regla nunca se llegarían a estimar valores iguales a 1. El modo de resolver este problema es tomar un umbral más pequeño. Ejemplo extraído de modelos de elección discreta (Medina, Eva).

Con cualquier tipo de regla predictiva similar se cometerán dos errores: habrá ceros que se clasifiquen incorrectamente como unos y unos que se clasifiquen incorrectamente como ceros. Si se reduce el umbral por debajo de 0.5 aumentará el número de veces que se clasifican correctamente observaciones para las que $Y_i = 1$, pero también aumentará el número de veces en que se clasifiquen observaciones como unos para las que $Y_i = 0$.

Cambiando el valor del umbral se reducirá siempre la probabilidad de un error de un tipo y se aumentará la probabilidad del otro tipo de error. Por lo que el valor que debe tomar el umbral depende de la distribución de datos en la muestra y de la importancia relativa de cada tipo de error.

Una vez seleccionado el nivel del umbral, y dado que los valores reales de Y_i son conocidos, basta con contabilizar el porcentaje de aciertos para decir si la bondad del ajuste es elevada o no. A partir de este recuento se puede construir el siguiente cuadro de clasificación:

Cuadro de clasificación de aciertos

		Valor real de Y_i	
		$Y_i = 0$	$Y_i = 1$
Predicción de $\hat{P}_i$	$\hat{P}_i < c$	P_{11}	P_{12}
	$\hat{P}_i > c$	P_{21}	P_{22}

Donde P_{11} y P_{22} corresponderán a predicciones correctas (valores 0 bien predichos en el primer caso y valores 1 bien predichos en el segundo caso), mientras que P_{12} y P_{21} corresponderán a predicciones erróneas (valores 1 mal predichos en el primer caso y valores 0 mal

predichos en el segundo caso). A partir de estos valores se pueden definir los índices que aparecen en el siguiente cuadro.

Índices para Medir la Bondad del Ajuste

Índice	Definición	Expresión
Tasa de aciertos	Cociente entre las predicciones correctas y el total de predicciones.	$\frac{P_{11}+P_{22}}{P_{11}+P_{12}+P_{21}+P_{22}}$
Tasa de errores	Cociente entre las predicciones incorrectas y el total de predicciones.	$\frac{P_{12}+P_{21}}{P_{11}+P_{12}+P_{21}+P_{22}}$
Especificidad	Proporción entre la frecuencia de valores 0 correctos y el total de valores 0 observados.	$\frac{P_{11}}{P_{11}+P_{12}}$
Susceptibilidad	Razón entre los valores 1 correctos y el total de valores 1 observados.	$\frac{P_{22}}{P_{21}+P_{22}}$
Tasa de falsos ceros	Proporción entre la frecuencia de valores 0 incorrectos y el total de valores 0.	$\frac{P_{21}}{P_{11}+P_{21}}$
Tasa de falsos unos	Razón entre los valores 1 incorrectos y el total de valores 1 observados.	$\frac{P_{12}}{P_{12}+P_{22}}$

Prueba de Hosmer – Lemeshow

Otra medida global de la exactitud predictiva, no basada en el valor de la función de verosimilitud sino en la predicción real de la variable dependiente, es el contraste de clasificación diseñado por Hosmer y Lemeshow (1989). Dicho contraste consiste en realizar comparaciones entre el valor estimado y el observado por grupos. Para ello las observaciones se dividen en J grupos (generalmente 10) aproximadamente iguales, dividiendo el recorrido de la probabilidad en deciles de riesgo (esto es probabilidad de ocurrencia del fenómeno < 0.1, < 0.2, y así hasta <1). Cada uno de los grupos contiene n_j observaciones, y en cada uno de los J grupos se define:

- Y_j Como la suma de valores 1 en cada uno de los grupos $(Y_j = \sum Y_i)$.

- $\bar{P}_j$ como la media de los valores predichos en cada grupo $\left(\bar{P}_j = \Sigma \frac{\hat{P}_i}{n_j}\right)$.

A partir de esta información se puede construir una tabla de contingencia a través de la que se compara tanto la distribución de ocurrencia, como la de no ocurrencia prevista por la ecuación y los valores realmente observados. El contraste se realiza comparándolas frecuencias observadas y esperadas a través del cálculo del estadístico:

$$HL = \sum_{j=1}^{J} \frac{(Y_j - n_j \bar{P}_j)^2}{n_j \bar{P}_j (1 - \bar{P}_j)}$$

Hosmer y Lemeshow demuestran que cuando el modelo es correcto el estadístico HL sigue una distribución chi-cuadrado con J-2 grados de libertad, por lo que valores inferiores del estadístico calculado respecto al teórico indicarán un buen ajuste del modelo.

El uso correcto de este contraste requiere un tamaño de muestra adecuado para asegurar que cada grupo cuenta al menos con cinco observaciones. Además el estadístico chi-cuadrado es sensible al tamaño muestral, permitiendo que esta medida encuentre diferencias estadísticamente muy pequeñas cuando el tamaño muestral crece.

Precisión en la Predicción

Un indicador de que tan bien se desempeña el modelo está en su habilidad para clasificar a los casos con precisión en las dos categorías que están definidas en la variable predicha, la precisión predictiva global y las precisiones especificas se obtienen en el cuadro de porcentajes a través de una división del número de casos estimados de forma correcta sobre el total de casos clasificados para ambas categorías conocidos antes de iniciar el trabajo de regresión logit. Estos valores son importantes para generar respuestas inmediatas en cuanto a la buena predicción que se obtiene con el modelo, y

debe ser comparada de inmediato con las pruebas de significancia para verificar si ambos objetivos son compatibles o las diferencias son notorias, bajo el concepto inicial de la falta de correspondencia entre el ajuste del modelo y los estadísticos de verosimilitud, o la significancia delas variables individuales y la habilidad predictiva del mismo. Dado que encontrar un modelo significativo no es razón suficiente de tener una elevada predictibilidad.

A los valores de porcentajes particulares para la tabla de aciertos se les conoce como especificidad que se hace en función de la categoría más común o específica y para la categoría menos común se le conoce como sensibilidad, por ser la categoría sensible o de interés sobre el estudio y utilización del modelo logit.

Realizando predicciones

Con los coeficientes de regresión podemos hacer predicciones sobre los valores de casos individuales, entonces vamos a calcular la probabilidad de la categoría de interés en el estudio, remplazando los valores apropiados individuales en cada variable para dicho caso individual en la función matemática del modelo ya conocido, este resultado se puede expresar en términos de oportunidad con la consideración exponencial que de elevar dicha probabilidad al valor (e) como se discutió para los valores de los coeficientes del modelo. Y la probabilidad por si sola sea utilizada para llevar dicho caso individual a una de las dos categorías de la variable dependiente.

Curva Operativa de Rendimiento (ROC)

Con la tabla de clasificación de aciertos, se describió su significado de los porcentajes encontrados para las dos categorías de la variable predicha, entonces todo parte del conocimiento que el punto de corte en el análisis logit es 0.5, porque así lo define por defecto el programa estadístico y la naturaleza

de las investigaciones lo determinan de forma general de esa manera. En la parte ultima de la regresión logit se obtiene la curva ROC, (por sus siglas en inglés) cuyos ejes son la sensibilidad o susceptibilidad (eje y) y el complemento de la especificidad (eje x) y al formar las coordenadas para la curva se obtiene los valores de referencia para el patrón de desempeño de las proporciones, y ver la relación que guarda con la variable de estado.

El uso inicial de esta curva es ver los balances entre sus dos ejes, es probable que se esté interesado en cambiar el punto de corte debido a la experiencia de los investigadores y la investigación empieza nuevamente con una percepción diferente, pero hay que tener en cuenta que esto genera un análisis previo de la distribución inicial de las categorías de estudio.

III. METODOLOGÍA

3.1. TIPO DE INVESTIGACIÓN

3.1.1. Según la naturaleza del objeto de estudio

Factual o empírica: Las ciencias naturales y las ciencias sociales tienen como objeto de estudio los hechos materiales, los fenómenos que son visibles en la realidad; por eso en estas ciencias se realiza investigación factual o empírica, es decir, investigación referida a los hechos observables en la realidad .

3.1.2. Según el tipo de pregunta planteada en el problema

Práctica: Plantean la modificación o la transformación de la realidad en los términos más convenientes para el hombre.

3.1.3. Según el método de estudio de las variables

Investigación Cuantitativa: La investigación cuantitativa se realiza cuando el investigador mide las variables y expresa los resultados de la medición en valores numéricos.

3.1.4. Según el número de variables

Multivariada: Las **investigaciones** multivariadas o factoriales consideran que el efecto es producido por la concurrencia de dos o más variables independiente que actúan sobre la variable dependiente.

3.1.5. Según el tipo de datos que producen

Primaria: Porque es una fuente directa donde se conseguirán los datos que será la información involucrada en la investigación; esto se refiere a que recopilamos la información a través de un instrumento de medición utilizado como parte de la metodología del trabajo.

3.1.6. Según el tiempo de aplicación de la variable

Transversales o sincrónicas: El tipo de investigación, de acuerdo a la finalidad es aplicada y de acuerdo a la técnica de contrastación, es correlacional explicativa. El diseño del presente trabajo de investigación es retrospectivo y de corte transversal.

3.2. POBLACIÓN DE ESTUDIO

La población a la que se desea generalizar los resultados obtenidos incluyó a las personas de 40 años a más de edad registradas en el listado del cuestionario del hogar de la ENDES aplicado en el departamento de Piura. Población que organiza INEI de acuerdo a criterios estadísticos. La población de estudio lógicamente está dirigida a las personas en general del departamento de Piura, con la característica única de la edad de los posibles encuestados en este estudio, el tamaño de población del departamento es un valor no conocido de forma integral, pero se conoce el número de encuestas que se realizan en aproximado en un año laboral, por parte del personal que participan en esta evaluación de salud, sabiendo que este número de encuestas es igual a 840 encuestas en aproximado considerando áreas urbanas y rurales que conforman el departamento en todas sus provincias y distritos respectivamente.

La población que se analizó fueron personas que voluntariamente aceptaron su participación. Gracias a la utilización de la metodología por conglomerados bietapicos, utilizadas en las evaluaciones del INEI, y es una ayuda importante,

dado que la investigación aportara, en lo que a mí respecta un valor agregado a mi conocimiento especial de este tipo de investigación, al ser parte del personal que labora actualmente en las oficinas de esta institución, participando de los diversos censos y encuestas de evaluación de programas sociales afines de mi proyecto de investigación. La población sabemos está sometida a las condiciones propias de las encuestas reguladas según el INEI y ciertos criterios individuales definidos por las oficinas de este establecimiento de gobierno.

3.3. SELECCIÓN DE MUESTRA

En el ámbito epidemiológico, en un estudio con regresión logística se utiliza la fórmula de Freeman [n = 30*(k +1)] o lo que es lo mismo, en términos generales unas treinta veces el número de variables independientes más uno, dado que se cuenta con aproximadamente 13 variables, llamados también factores de salud, la muestra será de 420 personas, seleccionadas adecuadamente según las características de selección antes expuestas, tomados aleatoriamente de los diferentes distritos, asentamientos humanos, y zonas rurales que comprenden el departamento de Piura, conociendo además que el tamaño de población del departamento de Piura es muy grande y en términos estadísticos lo asumimos como desconocido, entonces se trata de obviar este dato porque es un valor que supera las condiciones de recolección de datos factible en los recursos de esta evaluación. (Manual de Técnicas Estadísticas Multivariadas con SPSS, Regresión Logística Binaria; 2012 – pág. 49).

Debemos tener conocimiento que son dos modelos de regresión utilizados y el tamaño de muestra es el mismo en ambos casos dado que son los mismos factores implicados en ambas evaluaciones. La institución tiene un plan de muestreo ya descrito que a su vez está plenamente validado y es utilizado en todas las investigaciones a nivel nacional, por lo tanto describimos esta situación para conocimiento y teniendo claro que las personas elegidas que han sido encuestadas y que participan en este estudio son parte de la base de datos con la que se maneja para la propiciación de resultados de esta investigación, cada uno de esos

individuos están plenamente incluidos e identificados, en nuestra información y que será el punto de partida para el análisis estadístico que deseamos realizar.

3.4. METODO Y PROCEDIMIENTO

En este trabajo se utilizó una técnica estadística multivariada, conocida como Regresión Logística Binaria, que es un procedimiento de análisis ideal para evaluar este tipo de investigación, dado que las características de la investigación son las adecuadas para involucrar esta forma de trabajo; en relación al tipo de variables incluidas y sobre todo los objetivos principales y específicos que estamos desarrollando.

El tamaño de muestra ya está calculado de acuerdo a las formulas precisas que se definen en la teoría de la regresión logística, en base al número de variables con las que se contó en esta investigación, sin dejar de lado por supuesto los criterios de representatividad de la muestra, y el tamaño adecuado para cada lugar donde se encuestara, los criterios fueron usados de la forma más responsable, en virtud de los conocimientos y conceptos con los que se cuentan como parte de los fundamentos básicos que comprenden nuestra especialidad.

La información quedo registrada en las encuestas, luego se introdujo en el programa Excel y IBM SPSS 20, donde se generó una base de datos con todas las variables dependientes e independientes, y desde allí se realizó un análisis exploratorio para verificar los datos erróneos, perdidos y en blanco obtenidos al recolectar la información y se ejecutó los pasos necesarios para corregir estos posibles errores.

La información no necesita ser validada dado que el tipo de cuestionario que se utiliza dentro del diseño de la encuesta está debidamente validado por las reglas estadísticas que acreditan la confiabilidad de la encuesta, estas reglas son el alfa de crombach principalmente, y que se realiza desde la oficina de control de calidad de la información en la institución donde laboro, por lo tanto podemos

asumir que la recepción de la información evaluada en su conjunto con todas las circunstancias antes descritas, fue el inicio para introducir los datos al programa ideal para la evaluación, y empezar a construir el modelo y los resultados que se esperó encontrar.

El método que se utiliza dentro de la regresión logística es el conocido como método hacia adelante, que realiza la inclusión de las variables de forma individual una por una identificando cual es significativa y que va ser introducida dentro del modelo y descartando aquellas que no son significativas, el método por pasos se refiere también a una estructura por bloques de variables, acerca de aquellas que se van generando en cada inclusión realizada por la técnica estadística.

Las variables son todas aquellas incluidas dentro de la encuesta de trabajo de la investigación, y se pueden definir dentro del cuadro de resumen de todas variables, lo que debemos agregar es que la difusión de estas variables o componentes de la encuesta están definidas de acuerdo a un trabajo de evaluación muestral para estudios de salud incluidos en el INEI, en ellas se toma en cuenta variables Dicotómicas, en algunos casos variables con más de dos posibles respuestas, variables numéricas, que fueron redefinidas según las características de SPSS, para un trabajo más consolidado y de mejor comprensión teórica, y práctica.

3.5. DEFINICIÓN CONCEPTUAL Y OPERACIONAL DE VARIABLES

3.5.1. Variable Dependiente

Debemos tener en cuenta que la variable dependiente es una variable Dicotómica, es decir solo puede presentar dos posibles respuestas, y esta termina siendo ideal para la utilización del modelo de regresión logística binaria (Logit), además esta solo tiene una conclusión inmediata acerca de las respuestas de nuestros encuestados, donde se llevara a cabo la confirmación

de la existencia de las enfermedades no trasmisibles admitidas como fundamento de esta evaluación, es por eso que podremos considerar la utilización de dos modelos de trabajo para cada una de las enfermedades.

Diabetes o Azúcar Alta en la Sangre

Es el resultado de un desorden en el metabolismo de los alimentos en el cuerpo humano, la misma que se caracteriza por una elevada concentración de glucosa en la sangre; ya sea por falta de la hormona llamada insulina, segregada por el páncreas, que posibilita su absorción por las células del cuerpo humano; o, por que las células del cuerpo no responden al estímulo de la insulina generada por el páncreas (Manual de la entrevistadora - ENDES, 2013, pág.107)

Presión Alta o Hipertensión Arterial

Es la elevación persistente de la presión arterial por encima de los límites considerados como normales. La presión arterial es la fuerza con que la sangre empuja las paredes de los vasos sanguíneos (Manual de la entrevistadora - ENDES, 2013, pág.109).

3.5.2. Variable Independiente

Estas variables están relacionadas a las características personales y del comportamiento de la población que se sospecha constituye factores de riesgo para desarrollar la Hipertensión Arterial y Diabetes Mellitus. Estas variables que en la investigación son llamados factores independientes con sus respectivas valoraciones y definiciones son necesarias para su comprensión, dado que la interpretación de los resultados está basada en las condiciones particulares e individuales de cada una de ellas.

3.5.3. Operacionalización de las variables dependientes e independientes de la encuesta

ENFERMEDADES CRÓNICAS

01	¿Algún médico u otro profesional de la salud le ha dicho que usted tiene diabetes o azúcar alta en la sangre?	SI..... 1 NO..... 2 **(PASE A 03)**
02	Actualmente, ¿Usted recibe tratamiento médico para la diabetes?	SI..... 1 NO..... 2
03	¿Algún médico u otro profesional de la salud le ha dicho que usted tiene "Presión Alta" o Hipertensión Arterial?	SI..... 1 NO..... 2 **(PASE A 06)**
04	Actualmente, ¿Usted recibe tratamiento médico para la "Presión Alta"?	SI..... 1 NO..... 2
05	¿Hace cuánto tiempo que el médico u otro profesional de la salud le diagnosticaron que usted tiene "Presión Alta" o Hipertensión Arterial?	MESES..... 1 ☐☐ AÑOS..... 2 ☐☐
06	Normalmente, ¿Su actividad diaria la realiza de pie o sentado/a?	DE PIE..... 1 SENTADO/A..... 2
07	Normalmente, ¿Qué tanto esfuerzo físico le demanda a usted realizar su actividad diaria: leve, moderado o alto?	LEVE..... 1 MODERADO/ ALTO..... 2
08	Normalmente, ¿Usted practica algún deporte o realiza algún ejercicio físico como planchas, caminatas u otro similar al menos un día a la semana?	SI..... 1 NO..... 2
09	El día domingo, por lo normal ¿Cuántas horas usted ve televisión y/o películas en casa?	NO VE TV O VIDEOS..... 1 MENOS DE TRES HORAS... 2 TRES O MÁS HORAS..... 3

10	¿Suele usted agregarle sal a su plato de comida para atender su gusto personal?	SI..... 1 NO..... 2
11	Normalmente, ¿Usted acompaña con ensalada de verduras el consumo de menestras?	SI..... 1 NO..... 2
12	Normalmente, ¿Usted consume fruta fresca todos los días de la semana?	SI..... 1 NO..... 2
13	Normalmente, ¿Usted consume alguna golosina o postre entre las comidas diarias?	SI..... 1 NO..... 2
14	Normalmente, ¿Cuántos días de la semana come usted alguna fritura: papa frita, pollo frito, churrasco, pescado frito u otro similar?	N° DE DIAS..... ☐☐
15	Normalmente, ¿Usted come el pollo sin pellejo y/o la carne desgrasada?	SI..... 1 NO..... 2
16	¿Alguna vez usted ha fumado diariamente al menos un cigarrillo?	SI..... 1 NO..... 2 **(PASE A 20)**
17	¿A qué edad empezó a fumar diariamente al menos un cigarrillo?	EDAD EN AÑOS..... ☐☐
18	Actualmente, ¿Usted fuma diariamente al menos un cigarrillo?	SI..... 1 NO..... 2 **(PASE A 20)**
19	En promedio, ¿Cuántos cigarrillos fuma usted al día?	N° DE CIGARRILLOS AL DIA..... ☐☐
20	En el mes de (anterior) _____ ¿Ha tomado usted al menos un vaso de cachina, cerveza, vino, pisco, u otra bebida similar?	SI..... 1 NO..... 2 **(FINALIZAR)**
21	En ese mes ¿En cuántas ocasiones u oportunidades tomó usted?	N° DE OCASIONES QUE TOMÓ..... ☐☐
22	¿Cuántos vasos o botellas tomó usted la última vez?	N° DE VASOS.....1 ☐☐ N° DE BOTELLAS..... 2 ☐☐

23	¿Cuál de esas bebidas toma usted con mayor frecuencia?	CHICHA DE JORA... 01 CACHINA..... 02 CERVEZA..... 03 VINO..... 04 PISCO..... 05 RON..... 06 WHISKY..... 07 YONQUE / CAÑA... 08 MAZATO..... 09 OTRO_____ 10

3.6. RECOLECCIÓN DE DATOS

La recolección de los datos puede ser explicada de la siguiente manera en un contexto de pasos resumido:

La población objetivo con énfasis en el marco muestral que se debe tomar en cuenta son todos las personas de 40 a más años de edad de los distintos sectores, distritos, áreas rurales, asentamientos humanos del departamento de Piura, dado que contamos con la excelente colaboración de tener ya definido un muestreo por conglomerados bietapico plenamente estructurado y organizado de acuerdo a las características geográficas, y poblacionales de las evaluaciones generales por parte de la institución donde laboro actualmente (INEI).

IV. ANÁLISIS DE LOS RESULTADOS

4.1. CONSTRUCCIÓN DEL MODELO LOGIT

4.1.1. Especificación del Modelo de Regresión Logística

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \ldots + \beta_{13} X_{13}}}$$

Hipótesis:

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \cdots = \beta_{13} = 0$

$H_1: Al\ menos\ un\ \beta_i \neq 0, \forall\ i = 1, 2, 3, \ldots 12, 13$

H_0: Los factores de salud asociados al comportamiento de las personas (X_1, X_2, X_3, X_4, X_5, X_6,..., X_{13}) no influyen significativamente sobre la existencia de las enfermedades no transmisibles en la población de Piura (Y).

H_1: Al menos uno de los factores de salud asociados al comportamiento de las personas (X_1, X_2, X_3, X_4, X_5, X_6,..., X_{13}) si influyen significativamente la existencia de las enfermedades no transmisibles en la población de Piura (Y).

4.1.2. Ajuste del Modelo de Regresión Logística

Este ajuste ha sido efectuado mediante el software estadístico SPSS 20.0; se resumen los resultados a través de los cuadros mostrados. A partir de este momento se incluyen todos los cuadros necesarios, que tienen directa participación en el ajuste de ambos modelos que se están diseñando, puesto que se evalúan en la investigación dos enfermedades distintas. Por esta razón siempre se describirán dos cuadros en cada uno de los resultados que pertenecen al análisis de la regresión logística binaria y que arroja el programa estadístico, las interpretaciones son iguales en los dos casos, solo

cambia el valor numérico que se obtiene en cada cuadro para cada una de las enfermedades tratadas en la investigación realizada.

Cuadro 4.1: Casos incluidos en el análisis para la enfermedad Diabetes.

Resumen del procesamiento de los casos

Casos no ponderados		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	420	100,0
	Casos perdidos	0	,0
	Total	420	100,0
Casos no seleccionados		0	,0
Total		420	100,0

Cuadro 4.2: Casos incluidos en el análisis para la enfermedad Hipertensión.

Resumen del procesamiento de los casos

Casos no ponderados		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	420	100,0
	Casos perdidos	0	,0
	Total	420	100,0
Casos no seleccionados		0	,0
Total		420	100,0

Los cuadros de resumen del procesamiento de los casos dan a conocer que del total de la muestra seleccionada y procesada ningún dato ha sido excluido. Por lo tanto se debe proseguir con el análisis de resultados, al tener conocimiento que no hay datos perdidos, y no habrá problemas con respecto a ese supuesto necesario en la regresión Logit. Se debe tener en cuenta lo siguiente, dado que el número de factores implicados en el análisis es igual a 13 para ambas enfermedades es lógico suponer que el tamaño de la muestra será igual para ambos procesamientos, pero debemos manifestar que las muestras seleccionadas son distintas esto quiere decir que para la primera enfermedad llamada Diabetes se hizo una selección aleatoria del primer conjunto de datos, y después para la segunda enfermedad llamada Hipertensión se realizó una nueva selección aleatoria para ejecutar el análisis correspondiente. En ambos procedimientos el número de casos seleccionados

de la base de datos general que se tiene para la investigación fue de 420 personas.

Cuadro 4.3: Codificación de la variable dependiente para la enfermedad Diabetes.

Valor original	Valor interno
No	0
Si	1

Cuadro 4.4: Codificación de la variable dependiente para la enfermedad Hipertensión.

Valor original	Valor interno
No	0
Si	1

La codificación que por defecto efectúa el software estadístico SPSS no ha tenido que realizar cambios internos dado que en la matriz de datos la variable dependiente se ha encontrado codificada como: No (0), Si (1). Esto hace referencia a la existencia o no de las enfermedades no transmisibles (Diabetes e Hipertensión) en cada uno de los sujetos o personas incluidas en los dos conjuntos de datos seleccionados a partir de la base general con la que se cuenta, dichos conjuntos lógicamente vienen a ser nuestras muestras aleatorias.

Bloque 0: Bloque Inicial

IBM SPSS Statistics proporciona algunos resultados básicos bajo el título de "Bloque 0: Bloque Inicial". Estos se basan en un modelo logístico que contiene sólo un intercepto (constante). Aunque este modelo no es interesante, se obtiene alguna información básica y algunos resultados que son utilizados en las comparaciones o diferencias involucradas en las operaciones matemáticas de las pruebas estadísticas que intervienen en el modelo de regresión logística binaria.

Primero en las tablas de clasificación (Cuadro 4.5 y Cuadro 4.6), indica que en cada uno de los modelos siempre se predice la categoría más común (No: Ausencia de la enfermedad) y esto es correcto para el 96% y 84% del total de casos para las muestras seleccionadas en cada caso respectivamente. Este predice correctamente a todas las personas sin la enfermedad, pero ignora a todos los que si las presentan. Esto genera una línea base para lo que se debe evaluar posteriormente con la inclusión de todas las variables o factores dispuestos en esta investigación, los resultados porcentuales son los valores mínimos que deberíamos a estar dispuestos a aceptar al final en el porcentaje de aciertos globales con el modelo de regresión logística binaria.

Cuadro 4.5: Tabla de clasificación de los datos para la enfermedad Diabetes.

Observado			Pronosticado		
			Le han dicho que Ud. tiene diabetes o azúcar alta en la sangre		Porcentaje correcto
			No	Si	
Paso 0	Le han dicho que Ud. tiene diabetes o azúcar alta en la sangre	No	403	0	100,0
		Si	17	0	,0
	Porcentaje global				96,0

Cuadro 4.6: Tabla de clasificación de los datos para la enfermedad Hipertensión.

Observado			Pronosticado		
			Le han dicho que tiene Presión Alta o Hipertensión Arterial		Porcentaje correcto
			No	Si	
Paso 0	Le han dicho que tiene Presión Alta o Hipertensión Arterial	No	353	0	100,0
		Si	67	0	,0
	Porcentaje global				84,0

Las tablas adicionales en este bloque inicial presentan información sobre las variables en la ecuación (sólo la constante) y las que no están en ella. Estas se presentan en el Anexo de esta investigación. Estos cuadros no son de carácter importante es por ello, que solo se describen en la parte ultima del trabajo para referencia en algún momento de las interpretaciones de resultados que se tiene en los cuadros posteriores.

Bloque 1: Por pasos hacia adelante (Razón de Verosimilitudes)

Los algoritmos por pasos encuentran un subgrupo de variables que maximicen la verosimilitud. A continuación se muestran los cuadros incluidos en esta parte pero solo se tiene en cuenta el último paso implicado en el análisis, para una mejor comprensión.

Vemos que el análisis por pasos, con la selección hacia adelante y el estadístico razón de verosimilitud, tomó tres y cuatro pasos respectivamente en cada una de las enfermedades llamadas Diabetes e Hipertensión. Por eso sabemos que se seleccionaron solo tres y cuatro factores de salud asociados al comportamiento de las personas en cada uno de los modelos de regresión logística binaria encontrados, aunque en los cuadros implicados en esta parte no se identifica cuales factores ingresaron en la ecuación en cada paso, si se tiene el último paso en cada modelo estadístico con sus factores que resultaron como significativos. Las tablas de las pruebas ómnibus de los modelos por pasos, muestra el estadístico de significancia del predictor que ingresó en el último paso, en tanto que los resúmenes del modelo y del bloque nos permiten ver una prueba de los coeficientes del modelo en ese punto. No es de sorprender que el coeficiente en el último paso sea significativo, como en el modelo y estos a su vez sean iguales, esto siempre suele pasar en la regresión logística binaria.

Cuadro 4.7: Tabla de las pruebas ómnibus de los coeficientes del modelo para la Diabetes.

		Chi cuadrado	gl	Sig.
Paso 3	Paso	4,313	1	,038
	Bloque	26,863	3	,000
	Modelo	26,863	3	,000

La probabilidad de los resultados observados, dados los cálculos del parámetro, se conoce como verosimilitud. Por lo general se utiliza -2 veces el logaritmo natural de la verosimilitud (-2LL) como una medida del ajuste del modelo, dado que tiene vínculos con la distribución de Chi- cuadrado. Un

buen modelo que tiene una elevada verosimilitud se traduce en un valor pequeño de -2LL. En un ajuste perfecto, -2LL sería igual a 0.

En el cuadro N° 7.1 del anexo de la investigación, se observa que en el modelo que es analizado en el bloque inicial (bloque 0), el valor de -2LL es 142.342, este valor comparado con el obtenido en el paso tres del bloque 1, se reduce en 27 unidades aproximadamente. Es decir conforme ingresan variables adicionales en el modelo, la bondad de ajuste mejora, y eso se puede notar en el estadístico de verosimilitud -2LL, que disminuye.

Este cuadro, al igual que el cuadro a continuación, lo que se muestra y se resume es el estadístico Chi- cuadrado del modelo que es una prueba estadística de la hipótesis nula, acerca de que los coeficientes para todos los términos del modelo son igual a cero. Esto es equivalente definitivamente a la prueba F en la regresión lineal o también conocida como la tabla ANOVA, en esa parte de aquellos modelos estadísticos. Lo que se busca dentro de esta prueba ómnibus es verificar si el modelo es adecuado y esto se podrá comprobar con la significancia que se puede obtener al realizar la prueba. El valor del estadístico que obtenemos es igual 26.863 como se puede ver en el cuadro, este resultado que viene a ser la diferencia entre el -2LL inicial (un modelo solo con la constante) y el mismo coeficiente -2LL final (el modelo que incluye a todas las variables independientes). Tiene 3 grados de libertad, que representan la diferencia entre el número de parámetros en los dos modelos. Rechazamos la hipótesis nula porque la significancia es muy baja (0.000) y concluimos que el grupo de variables mejora la predicción del logaritmo natural de las oportunidades.

Cuadro 4.8: Tabla de las pruebas ómnibus de los coeficientes del modelo para la Hipertensión.

		Chi cuadrado	gl	Sig.
Paso 4	Paso	4,092	1	,043
	Bloque	61,953	4	,000
	Modelo	61,953	4	,000

Este cuadro 4.8, tiene la misma interpretación que el cuadro anterior, lo que se muestra y se resume es el estadístico Chi- cuadrado del modelo que es una prueba estadística de la hipótesis nula, acerca de que los coeficientes para todos los términos del modelo de la segunda enfermedad llamada Hipertensión son igual a cero. El valor del estadístico que obtenemos ahora es igual 61.953 como se puede ver en el cuadro, que viene a ser la diferencia entre el -2LL inicial (un modelo solo con la constante) y el mismo coeficiente -2LL final (el modelo que incluye a todas las variables independientes). Tiene 4 grados de libertad, que representan la diferencia entre el número de parámetros en los dos modelos. Rechazamos la hipótesis nula porque la significancia es muy baja (0.000) y concluimos que el grupo de variables mejora la predicción del logaritmo natural de las oportunidades. El valor inicial del -2LL según lo que se puede ver en el anexo es igual a 368.659 lo cual asume una diferencia de 62 unidades aproximadamente con respecto al valor final, eso lo demuestra el resultado que se ha obtenido en el cuadro 4.7, que en definitiva es lo que se debe comprobar siempre en los cuadros de las pruebas ómnibus de los coeficientes del modelo, es decir la conclusión es la misma, cuando ingresan variables al modelo el coeficiente de verosimilitud disminuye, lo cual significa una mejora en la bondad de ajuste del modelo de regresión logística binaria. A continuación en los siguientes cuadros se puede obtener también el estadístico de verosimilitud final que se han usado en estas comparaciones de diferencia.

Cuadro 4.9: Tabla del resumen del modelo y análisis de los coeficientes de determinación para la enfermedad Diabetes.

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
3	115,479[a]	,062	,216

Ahora podemos notar el estadístico de verosimilitud final para el primer modelo de la enfermedad llamada Diabetes, existe una relación entre el análisis de los cuadros anteriores puesto que se presentan valores en común, con todas las variables del modelo la bondad de ajuste del estadístico -2LL es 115.479, que se presenta en la tabla. También se muestran los valores que son

análogos a la R cuadrado en la regresión estándar, pero dada la relación funcional que existe entre la media y la desviación estándar de la variable dependiente en el modelo Logit, por ser una variable dicotómica, la cantidad de varianza explicada por el modelo se debe definir diferente, la R cuadrado de Cox y Snell es igual 0.062 y la R cuadrado de Nagelkerke es igual a 0.216 por lo general se prefiere esta última sobre la primera porque puede llegar a tomar un valor máximo de 1. A través de cualquiera de ambos valores se puede ver, que solo el modelo explica una cantidad mínima de la varianza total. En el método por pasos los estadísticos de determinación suelen aumentar según el número de pasos que se van incluyendo en el análisis. Con los resultados presentados podemos notar que las variables incluidas no son determinantes para explicar la variación o dispersión de los errores, eso se puede interpretar en la misma forma como se hace en el análisis de regresión lineal, la conclusión inmediata que podemos obtener es que aunque el modelo de la enfermedad Diabetes tiene un buen ajuste existen otras variables o factores que deberían ser consideradas en un estudio posterior que ayude a explicar de mejor manera la varianza total del modelo de regresión logística binaria.

Cuadro 4.10: Tabla del resumen del modelo y análisis de los coeficientes de determinación para la Hipertensión.

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
4	306,705[a]	,137	,235

De la misma forma se hace la interpretación para el modelo de la segunda enfermedad llamada Hipertensión en esta investigación, con todas las variables del modelo la bondad de ajuste del estadístico -2LL es 306.705, que se presenta en la tabla. La cantidad de varianza explicada por el modelo ha obtenido los siguientes valores, la R cuadrado de Cox y Snell es igual 0.137 y la R cuadrado de Nagelkerke es igual a 0.235 por lo general se prefiere esta última sobre la primera porque puede llegar a tomar un valor máximo de 1, como se puede notar una vez más los valores encontrados en el cuarto paso son también bastante mínimos en relación con la cantidad de varianza que

pueden llegar a explicar, tienen una leve diferencia de aumento en comparación con el primer modelo, pero sigue siendo un valor pequeño en consideración con el porcentaje total de determinación que se desearía encontrar.

Cuadro 4.11: Tabla de la prueba de bondad de ajuste de Hosmer y Lemeshow para el modelo de la enfermedad Diabetes.

Paso	Chi cuadrado	gl	Sig.
3	,417	2	,812

Cuadro 4.12: Tabla de contingencias para la prueba de Hosmer y Lemeshow para el modelo de la enfermedad Diabetes

		Le han dicho que Ud. tiene diabetes o azúcar alta en la sangre = No		Le han dicho que Ud. tiene diabetes o azúcar alta en la sangre = Si		
		Observado	Esperado	Observado	Esperado	Total
Paso 3	1	262	262,616	4	3,384	266
	2	84	83,384	3	3,616	87
	3	38	37,384	3	3,616	41
	4	19	19,616	7	6,384	26

En esta parte de resultados aparecen dos tablas incluidas en un mismo análisis, que son los resúmenes de la prueba de bondad de ajuste de Hosmer y Lemeshow, la segunda tabla se le llama tabla de contingencia para la prueba estadística, esta se calcula dividiendo los casos en diez grupos de tamaño aproximadamente igual a partir de las probabilidades calculadas con el modelo encontrado, y luego comparando el número de valores observados con los esperados o predichos, en cada categoría de la variable dependiente.

Como se explica la teoría se debe tener muestras bastantes grandes para que esta prueba no tenga problemas al realizarla, pero si se observa la tabla de contingencia en el método por pasos se ejecuta una corrección de continuidad dentro de la prueba dado que los resultados para la segunda categoría de la variable dependiente son muy pequeños, este trabajo es algo similar a la prueba de bondad de ajuste Chi- cuadrado en tablas de contingencia

perteneciente a contrastes no paramétricos en estadística inferencial. Según la tabla no aparece una distribución de diez grupos como se espera, sino por el contrario se tiene una división de solo cuatro grupos como se observa en la tabla esta situación casi siempre ocurre cuando se tiene esta dificultad en la cantidad de datos para alguna de las categorías de la variable dependiente.

Los resultados de la prueba de bondad de ajuste de Hosmer y Lemeshow se muestran en el cuadro 4.11, en el cual se observa que la bondad de ajuste en el paso 3 es 0.417, y se distribuye como un valor de Chi-cuadrada con una significancia de 0.812. Al comparar los eventos observados con los esperados en el contexto de evaluar la bondad de ajuste, se espera encontrar una probabilidad no significativa o sea mayor al 5%, lo que indica que los eventos esperados y observados están cerca, lo que implica que el modelo tiene un buen ajuste. En este caso el modelo tiene un buen ajuste, lo que confirma el cambio en la prueba -2LL (prueba del modelo). La prueba de ajuste de Hosmer y Lemeshow es no significativa en el paso tres como se observa en la tabla lo que representa un signo promisorio, lo cual nos daría a conocer de forma previa un buen modelo de regresión para realizar predicciones, pero debemos ser prudentes debido al tema de las frecuencias muy bajas, de modo que la tabla que se utilizó para la prueba es muy pequeña, lo ideal siempre es tener diez grupos donde los casos estén colocados en base a los valores predichos con el modelo encontrado.

A continuación logramos obtener para la segunda enfermedad llamada Hipertensión los cuadros que le corresponden para esta prueba estadística donde se va a observar resultados similares y entonces las interpretaciones tienen que guardar las consideraciones de prudencia para no cometer errores y verificar al final del análisis la capacidad predictiva de estos modelos.

Cuadro 4.13: Tabla de la prueba de bondad de ajuste de Hosmer y Lemeshow para el modelo de la enfermedad Hipertensión.

Paso	Chi cuadrado	gl	Sig.
4	,300	1	,584

Cuadro 4.14: Tabla de contingencias para la prueba de Hosmer y Lemeshow para el modelo de la enfermedad Hipertensión.

		Le han dicho que tiene Presión Alta o Hipertensión Arterial = No		Le han dicho que tiene Presión Alta o Hipertensión Arterial = Si		
		Observado	Esperado	Observado	Esperado	Total
Paso 4	1	195	196,653	19	17,347	214
	2	122	120,587	17	18,413	139
	3	36	35,761	31	31,239	67

Los resultados de la prueba de bondad de ajuste de Hosmer y Lemeshow se muestran en el cuadro 4.13, para este segundo modelo a revisar tiene la misma representación conceptual; en el cual se observa que la bondad de ajuste en el paso 4 es 0.300, y se distribuye como un valor de Chi-cuadrada con una significancia de 0.584. Al comparar los eventos observados con los esperados en el contexto de evaluar la bondad de ajuste, nuevamente se encuentra una probabilidad no significativa, lo que indica que los eventos esperados y observados están cerca, lo que implica que el modelo tiene un buen ajuste. En este caso el modelo tiene un buen ajuste, lo que confirma el cambio en la prueba -2LL (prueba del modelo). Casi siempre obtener un buen ajuste en la prueba de Hosmer y Lemeshow nos da a entender que los cambios sucedidos en la función de verosimilitud han sido suficientes para confirmar que los modelos de regresión encontrados son adecuados para proseguir con el análisis.

Se debe considerar que en ambos modelos los grados de libertad de la prueba vienen representados como el número de categoría finales de las tablas de contingencia menos dos (K-2). Esto significa que las tablas ya reducidas a través de la corrección por continuidad, se consideran para saber el número de filas y a partir de ahí saber los grados de libertad implicados en la prueba de bondad de ajuste, en los casos de ambos modelos podemos notar que el número de filas en las tablas de contingencia fueron igual a cuatro y tres respectivamente, por lo tanto si realizamos la diferencia los grados de libertad resultantes serán igual a dos y un grado de libertad respectivamente y lo

podemos observar en las tablas presentadas de la prueba estadística perteneciente a estos modelos de regresión logística binaria.

Cuadro 4.15: Tabla de las variables incluidas dentro de la ecuación del modelo para la enfermedad Diabetes.

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
								Inferior	Superior
Paso 3[a]	Presencia_hipertension_Alta	2,030	,530	14,672	1	,000	7,610	2,694	21,499
	Nivel_Esfuerzo_Fisico	-1,204	,530	5,152	1	,023	,300	,106	,849
	Consume_Frutas	-1,100	,526	4,369	1	,037	,333	,119	,934
	Constante	-,181	1,187	,023	1	,879	,834		

En el cuadro 4.15 se muestra el valor de los coeficientes del modelo, obtenido en el tercer paso de este método para la primera enfermedad llamada Diabetes y el cual quedara como modelo final de presentación, esta última salida debe verse tan igual como la interpretación de la regresión lineal, se presentan la columna B y los errores estándar de los coeficientes B, también, los valores de la prueba basada en el estadístico de Wald y su nivel de significancia, como también la columna en la que se presentan los Odds ratio de cada variable, Exp (B). Finalmente se puede observar que el modelo se ha ajustado con un total de 3 variables de las 14 iniciales, se puede observar que todas ellas son significativas al nivel del 5%.

Para interpretar de forma sencilla, se debe recordar que el modelo original esta en términos del logaritmo natural de las oportunidades o logit. Por lo tanto, el coeficiente B es el efecto de una unidad de cambio en una variable independiente sobre el logaritmo natural de las oportunidades. El significado real en términos de la probabilidad que es lo que nos interesa llegar de forma más intuitiva se deduce en la columna Exp (B). Este se expresa ahora en términos de la razón de oportunidad sobre la variable dependiente según las categorías que pertenecen a cada variable independiente y de acuerdo a la forma como están establecidas dichas categorías en la encuesta.

Los factores que incrementan la probabilidad de la existencia de la enfermedad diabetes, según la descripción anterior es solo 1: presencia de Hipertensión en las personas. Por otro lado los factores asociados en el comportamiento de las personas que disminuyen la probabilidad de la existencia de la enfermedad diabetes son 2: Nivel de esfuerzo Físico, Consumo de frutas, estas dos últimas variables presentan un efecto negativo según el cuadro mostrado y esto se debe a que la categorías de los factores independientes son inversas a la descripción de la variable dependiente.

Cuadro 4.16: Tabla de las variables incluidas dentro de la ecuación del modelo para la enfermedad Hipertensión.

								I.C. 95% para EXP(B)	
		B	E.T.	Wald	gl	Sig.	Exp(B)	Inferior	Superior
Paso 4[a]	Presencia Diabetes	1,810	,580	9,726	1	,002	6,111	1,959	19,059
	Presencia Colesterol	-2,103	,351	35,964	1	,000	,122	,061	,243
	Realizacion_Deporte	,599	,297	4,088	1	,043	1,821	1,019	3,256
	Numero_horas_television	-,819	,271	9,129	1	,003	,441	,259	,750
	Constante	2,611	,995	6,885	1	,009	13,613		

En el cuadro 4.16 se muestra el valor de los coeficientes del modelo, obtenido en el cuarto paso de este método y el cual quedara como modelo final de presentación, este nuevo cuadro es de la misma forma que el anterior, para la segunda enfermedad llamada Hipertensión, una vez más se describe esta última salida y debe verse tan igual como la interpretación de la regresión lineal, sabemos que se presentan la columna B y los errores estándar de los coeficientes B, también, los valores de una prueba basada en el estadístico de Wald y su nivel de significancia, como también la columna en la que se presentan los Odds ratio de cada variable, Exp (B). Finalmente se puede observar que el modelo se ha ajustado con un total de 4 variables de las 14 iniciales, se puede observar que todas ellas son significativas al nivel del 5%.

Los factores asociados al comportamiento de las personas que incrementan la probabilidad de la existencia de la enfermedad Hipertensión, según la descripción anterior son solo 2: presencia de Diabetes en las personas y además realización o practica algún deporte por parte de las personas

encuestadas. Por otro lado los factores que disminuyen la probabilidad de la existencia de la enfermedad Hipertensión son también 2: Presencia de Colesterol y número de horas en ver televisión, estas dos últimas variables presentan un efecto negativo según el cuadro mostrado.

La interpretación de cada una de las variables incluidas en ambos modelos tiene que ver estrictamente con la forma de codificación como viene ya estructurada el tipo de encuesta utilizada por la institución de donde provienen los datos, y los objetivos que la institución persigue dentro del estudio de salud, para la comprensión clara de esta situación, se trata de concluir más adelante de forma precisa la interpretación individual de cada variable en su respectivo modelo y además de establecer la interpretación del resultado llamado odds ratio en cada factor encontrado específicamente en ambos modelos de regresión.

4.2. MODELO FINAL Y ODDS RATIO PARA CADA FACTOR

De acuerdo a los resultados del modelamiento con regresión logística en SPSS mediante el método "Adelante" con el criterio de Razón de Verosimilitudes se ha obtenido los siguiente modelos, puesto que esta investigación comprende de dos enfermedades se expresara dichas ecuaciones finales como sigue, teniendo ambas una forma general y similar pero que solo cambia por la clasificación de variables independientes incluidas en cada uno de ellos, se presentan a continuación:

Ecuación 1: expresión de la ecuación del modelo para la Diabetes.

$$p = \frac{1}{1 + e^{-Z}}$$

Dónde:

$$Z = -0.181 + 2.030 * \text{Pres\_Hp} - 1.204 * \text{Niv\_Esf} - 1.100\ \text{Cons\_Frut}$$

Ecuación 2: expresión de la ecuación del modelo para la Hipertensión.

$$p = \frac{1}{1 + e^{-z}}$$

Dónde:

$$Z = 2.611 + 1.810 * Pres_{Diab} + 0.599 * Real_{Dep} - 2.103\ Pres_{Col} - 0.819\ Num\_HTel$$

Del cual sus coeficientes β_i son estadísticamente significativos al nivel del 5%. El modelo se ajusta correctamente de acuerdo a la prueba de Bondad de Ajuste de Hosmer y Lemeshow, realizada anteriormente en ambos casos.

Ambas ecuaciones son las que se utilizaran para el cálculo de probabilidades y que son necesarias para encontrar el porcentaje de aciertos del modelo que se realiza a continuación.

En lo que respecta al cálculo de los odds ratio de cada factor significativo debemos tener en cuenta que ya se consideró su explicación en la parte anterior a estos resultados, el odds ratio se encuentra al hallar la siguiente expresión Exp (β_i) teniendo en cuenta que solo nos interesa tener el valor de esta expresión matemática; para los factores considerados en esta investigación que resultaron significativos, como se ha establecido y se puede observar en la tabla de variables incluidas en los modelos de regresión logística binaria, dicha operación matemática nos ofrece un resultado mayor a cero siempre y que debe ser interpretado de acuerdo a las categorías de la variable o factor al cual pertenece en comparación o relación con la variable dependiente de esta investigación.

Para la primera enfermedad llamada Diabetes los factores que resultaron significativos fueron los siguientes con su respectivo odds ratio calculado a partir del valor del coeficiente estimado para la ecuación resultante:

Presencia de Hipertensión	7.610
Nivel esfuerzo físico	0.300
Consumo de frutas	0.333

Las personas encuestadas que si presentan hipertensión que coincidentemente viene a ser la otra enfermedad involucrada, en este estudio tienen 7.6 veces más oportunidad de desarrollar Diabetes Mellitus a comparación de aquellos que no presentan dicha enfermedad, los dos factores siguientes sabemos que su coeficiente era negativo en consecuencia el valor del indicador calculado es inferior a la unidad, según la teoría nos dice que cuando el valor es menor que la unidad y por lo tanto es un numero decimal entre cero y uno, el factor o factores con esa característica no son factores de riesgo sino más bien factores de protección de las personas ante la existencia de desarrollar la enfermedad, además tenemos las categorías de cada factor que están clasificadas de manera distinta a la variable dependiente y lo podemos notar en la metodología en la parte de la encuesta y Operacionalización de las variables, la interpretación del resultado numérico es hacerlo a través del antilogaritmo del coeficiente y una forma más idónea es hacerlo en forma de porcentaje aproximado y se puede decir que aquellas personas que realizan un esfuerzo físico moderado o alto tienen 3.33 veces menor probabilidad que se encuentran dentro de la categoría que pertenece a la existencia o desarrollo de la Diabetes a diferencia de aquellos que solo realizan un esfuerzo leve. En el caso del otro factor llamado consumo de frutas, las personas que no consumen tienen 3 veces menos oportunidad en la probabilidad calculada para la existencia de la Diabetes en comparación con aquellos que si consumen.

Para la segunda enfermedad que participa de nuestra investigación llamada hipertensión al igual que la justificación anterior se presentan cuatro factores significativos y se presentan a continuación con sus respectivos indicadores odds ratio:

Presencia de Diabetes	6.111
Presencia de colesterol	0.122

Realización de deporte	1.821
Número de horas en televisión	0.441

Como podemos observar existen dos factores significativos donde el indicador de odds ratio es positivo en el primer caso es para el factor denotado como presencia de diabetes en las personas encuestadas, en este caso el resultado numérico es igual a 6.111 lo cual indica que existe una oportunidad de seis veces más de desarrollar hipertensión a aquellas personas que si presentan diabetes ya confirmado en comparación con aquellos que no tienen. El otro factor denotado como realización de deporte su valor numérico conseguido es igual a 1.821 lo cual lo cual nos dice que las personas que no practican algún tipo de deporte presentan una oportunidad de 1.8 veces en desarrollar hipertensión en comparación con aquellas personas que si realizan actividad física en algún deporte. A su vez podemos asegurar que existen dos factores significativos en modelo de regresión logística donde el resultado es un número decimal inferior a la unidad, y conocemos y recalcamos una vez más que a dichos factores se les llama factores de protección de las personas ante la existencia de desarrollar la enfermedad y el primer factor encontrado llamado presencia de Colesterol en las personas encuestadas su indicador es igual 0.122 lo cual indica en aproximado que las personas que no presentan colesterol tienen 8.2 veces menos oportunidad en probabilidad de contraer Hipertensión a diferencia de aquellos que si presentan dicha enfermedad. De la misma manera de interpretación para el factor denominado número de horas en ver televisión el resultado es igual 0.441 en este caso en aproximado se puede decir que las personas que ven televisión más de 3 horas el día domingo tienen 2.27 veces menos oportunidad en la probabilidad de contraer hipertensión arterial en comparación con aquellas personas que están clasificadas en las dos categorías restantes.

Estas interpretaciones se logran de manera directa para el caso de odds ratio superior a la unidad mientras que para cuando sucede lo contrario se realiza cálculos previos con el modelo para definir el valor porcentual de las posibilidades de encontrar personas con la enfermedad según las categorías de análisis en el factor, teniendo en cuenta la disposición original de las categorías en

las variables independientes, llamadas factores para poder definir de manera precisa la interpretación del indicador odds ratio calculado, según el programa estadístico para ambos modelos de regresión logística binaria. El intervalo de confianza nos ayuda también a saber cuándo un factor es de riesgo o de protección, en ninguna de las posibles situaciones descritas se encontrará la unidad, entonces cuando si el intervalo supera la unidad el factor siempre se considerara como un factor de riesgo, mientras que si el intervalo es inferior a la unidad el factor será entonces siempre un factor de protección.

4.3. ESTIMACIÓN DEL PORCENTAJE DE PERSONAS ENCUESTADAS CLASIFICADAS CORRECTAMENTE CON EL MODELO LOGIT.

La tabla de clasificación (Cuadro 4.17), paso 3, muestra que el porcentaje global de predicciones correctas es 91.40% en el modelo, para la primera enfermedad estudiada llamada Diabetes; se observa también que el número de pacientes que presentan la enfermedad son pronosticados correctamente en 52.9% (sensibilidad) y del total de personas que no tiene la enfermedad de estudio el 93.1% (especificidad) se pronostican correctamente. Es de tener en cuenta que el porcentaje global es altamente influenciado siempre por la sensibilidad del modelo. Entonces cuando se consigue una sensibilidad muy alta la capacidad predictiva del modelo será mucho mejor. Si se trata de extender una comparación acerca de estos resultados, se sabe entonces que se considera a aquellas personas expuestas a la enfermedad, a quienes respondieron que si dentro de las encuestas utilizadas, aquel porcentaje es de 4.08% del total en la muestra seleccionada. Con el modelo logit se encuentra la sensibilidad del modelo y se llama así dado que viene sujeto a la categoría de interés dentro de la investigación. Este valor se calcula de forma simple al igual que para la categoría (Especificidad); en otras palabras es el resultado de los personas correctamente clasificados por el modelo sobre el número de personas para cada categoría en la condición original de los datos en la variable dependiente sin haber empleado la herramienta estadística,
Los valores son como siguen:

Sensibilidad: 9/17=52.9 Especificidad: 375/403= 93.1

Cuadro 4.17: tabla de clasificación de las predicciones del modelo para la enfermedad Diabetes.

Observado			Pronosticado		
			Le han dicho que Ud. tiene diabetes o azúcar alta en la sangre		Porcentaje correcto
			No	Si	
Paso 3	Le han dicho que Ud. tiene diabetes o azúcar alta en la sangre	No	375	28	93,1
		Si	8	9	52,9
	Porcentaje global				91,4

La tabla de clasificación (Cuadro 4.18), paso 4, muestra que el porcentaje global de predicciones correctas es 84.3% en el modelo, para la segunda enfermedad estudiada llamada Hipertensión; se observa también que el número de pacientes que presentan la enfermedad son pronosticados correctamente en 44.8% (sensibilidad) y del total de personas que no tiene la enfermedad de estudio el 91.8% (especificidad) se pronostican correctamente. Para este modelo de la segunda enfermedad en la base de datos sin procesar se sabe que el porcentaje de personas que tienen o presentan la enfermedad es igual a 15.95% del total de la muestra.

Los porcentajes globales de predicciones correctas, es un resultado que nos da a conocer que tan bueno es el modelo en función a las predicciones que llegue a realizar en el futuro, se dice por lo general que cuando el modelo supera el 90%, entonces se considera como un buen modelo, las conclusiones que podemos obtener en esta investigación se dan en función con esos resultados que muestran las tablas, pero debemos tener prudencia en su interpretación dado que los casos que están en la categoría de interés de la variable dependiente son pocos o mínimos en comparación con la muestra total.

Sensibilidad: 30/67=44.8 Especificidad: 324/353= 91.8

Cuadro 4.18: tabla de clasificación de las predicciones del modelo para la enfermedad Hipertensión.

			Pronosticado		
			Le han dicho que tiene Presión Alta o Hipertensión Arterial		Porcentaje
	Observado		No	Si	correcto
Paso 4	Le han dicho que tiene Presión Alta o Hipertensión Arterial	No	324	29	91,8
		Si	37	30	44,8
	Porcentaje global				84,3

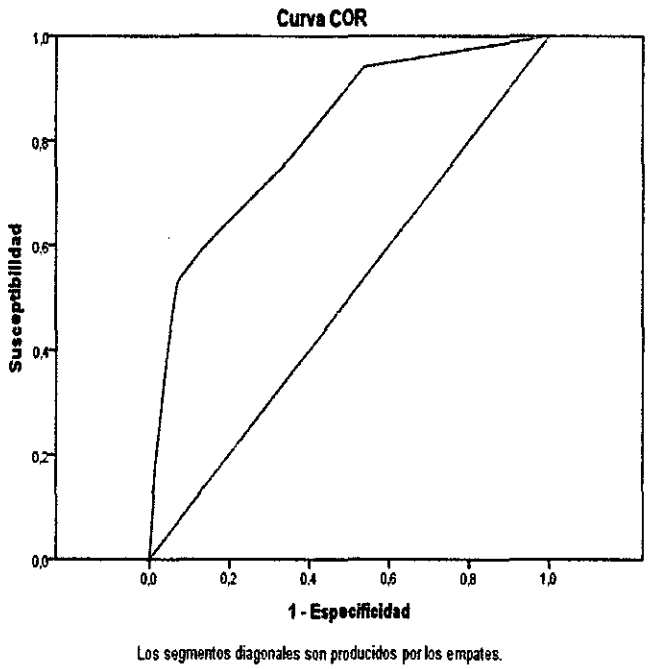
En ambas situaciones en las tablas de clasificación se puede notar que los valores porcentuales mínimos que estábamos dispuestos a aceptar para la capacidad de predicción correcta son casi iguales según la comparación que podemos hacer con los cuadros del bloque 0 : bloque inicial de la parte principal de este capítulo (Diabetes fue igual a 96% e Hipertensión fue igual a 84% y si observamos las dos últimas tablas los porcentajes encontrados son para Diabetes igual a 91.4% y para Hipertensión 84.3%) lo cual da a conocer que no ha surgido mejoras en cuanto a la clasificación utilizando el modelo de regresión logística binaria, por lo tanto se acepta los resultados en función de su valor, pero se debe considerar que los modelos no producen garantías en la predicción.

4.4. Estimación del valor de corte óptimo: Curva COR (Curva Operativa de Rendimiento)

En términos básicos, el procedimiento ordena los datos de acuerdo con la probabilidad predicha (puntuación) del evento de interés, luego calcula la sensibilidad y especificidad de cada cambio en la probabilidad predicha. Se puede examinar la gráfica o la tabla para evaluar los efectos de usar puntos de corte diferentes y los balances entre los tipos de éxitos y errores. La revisión de esta tabla o de la gráfica permite entender las ramificaciones de cambiar el punto de corte y ayuda a evaluar si uno diferente de 0.5 (especificado por defecto) satisface mejor las necesidades de predicción. En otros casos la comparación se realiza con el punto de corte elegido según las características de investigación que se está llevando a cabo.

Dado que lo ideal es tener una Sensibilidad y Especificidad igual a 1.0, entonces mientras más alejada se encuentra la curva COR de la diagonal principal, el método de diagnóstico es mejor. Como se observa en el gráfico 01 de esta investigación la Curva COR está regularmente alejada de la diagonal principal, por lo que se puede precisar una buena calidad predictiva de nuestro modelo de regresión logística binaria.

Gráfico 01. Curva Operativa de Rendimiento para la Diabetes.



Los segmentos diagonales son producidos por los empates.

En el cuadro 4.19, se aprecia que el área bajo la curva es 0.814 la cual difiere en mucho de 0.1 que sería el mínimo exigible para un método de diagnóstico, porque sabemos que en esta investigación para la enfermedad llamada Diabetes el punto de corte elegido es 0.1 en función con la teoría que nos da a conocer que ese valor es la probabilidad de que una persona en nuestra región tenga dicha enfermedad. El error estándar para esa estimación es 0.054 que multiplicado por 1.96 (para una confianza del 95.0%) y sumando y restando nos da el intervalo de confianza con un límite inferior de 0.707 y límite superior 0.920. De acuerdo a los resultados obtenidos se puede afirmar que el área bajo la Curva COR es significantemente mayor que lo mínimo exigible 0.1.

Además en el cuadro siguiente se puede observar que el punto de corte 0.068235 es el resultado de capacidad predictiva que proporciona un resultado distinto algo mejor en función de la categoría de interés, dado que se obtiene una Sensibilidad de 58.8% pero una reducción en la Especificidad con un valor de 87.3%. Esto hace suponer que se necesita un modelo de comparación similar para establecer conclusiones, cuando la sensibilidad del modelo aumenta se puede relacionar que la capacidad predictiva global del modelo también mejorara, es entonces de suponer que dicho cambio generara un resultado favorable en un modelo similar en el futuro.

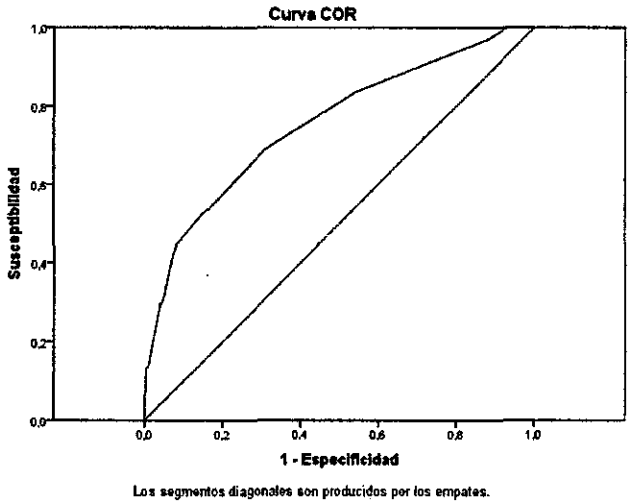
Cuadro 4.19: Área bajo la curva para la enfermedad Diabetes.

Variables resultado de contraste: Probabilidad pronosticada				
			Intervalo de confianza asintótico al 95%	
Área	Error típ.	Sig. asintótica	Límite inferior	Límite superior
,814	,054	,000	,707	,920

Cuadro 4.20: Coordenadas de la curva para la enfermedad Diabetes.

Variables resultado de contraste: Probabilidad pronosticada		
Positivo si es mayor o igual que	Sensibilidad	1 - Especificidad
,0000000	1,000	1,000
,0163249	,941	,536
,0256875	,765	,350
,0432682	,647	,199
,0682352	**,588**	**,127**
,1183992	,529	,069
,1670639	,412	,047
,2811426	,176	,015
1,0000000	,000	,000

Gráfico 02. Curva Operativa de Rendimiento para la Hipertensión



Ahora en los cuadros 4.21 y 4.22, se aprecia para la segunda enfermedad que el área bajo la curva es 0.755 la cual difiere en mucho de 0.3 que sería el mínimo exigible para un método de diagnóstico, dado que este fue el punto de corte elegido en esta enfermedad llamada Hipertensión. El error estándar para esa estimación es 0.035 que multiplicado por 1.96 (para una confianza del 95.0%) y sumando y restando nos da el intervalo de confianza con un límite inferior de 0.687 y límite superior 0.823. De acuerdo a los resultados obtenidos se puede afirmar que el área bajo la Curva COR es significantemente mayor que lo mínimo exigible 0.3.

Además en el cuadro siguiente se puede observar que el punto de corte 0.217475 es el resultado de capacidad predictiva que proporciona un resultado distinto algo mejor en función de la categoría de interés, dado que se obtiene una Sensibilidad de 52.2% pero una reducción en la Especificidad con un valor de 85.6%. Esto hace suponer que se necesita un modelo de comparación similar para establecer conclusiones, la predicción global para ambos modelos sugiere que se necesita un punto de corte distinto al seleccionado en base a la teoría, sin embargo este es un trabajo complejo donde debe participar la opinión de expertos en el tema, tanto en la parte teórica de la enfermedad como la función aplicativa de la técnica estadística de regresión y todo cuanto concierne su análisis.

Cuadro 4.21: Área bajo la curva para la enfermedad Hipertensión.

Variables resultado de contraste: Probabilidad pronosticada				
			Intervalo de confianza asintótico al 95%	
Área	Error típ.ª	Sig. asintótica	Límite inferior	Límite superior
,755	,035	,000	,687	,823

Cuadro 4.22: Coordenadas de la curva para la enfermedad Hipertensión.

Variables resultado de contraste: Probabilidad pronosticada		
Positivo si es mayor o igual que	Sensibilidad	1 - Especificidad
,0000000	1,000	1,000
,0426449	1,000	,929
,0608069	,970	,881
,0913856	,836	,541
,1279187	,687	,306
,1511776	,537	,164
,1841801	,537	,161
,2174753	**,522**	**,144**
,2669826	,448	,082
,3130815	,433	,076
,3457900	,418	,071
,4074501	,313	,048
,4716322	,299	,042
,5081524	,299	,040
,5445432	,134	,008
,6081157	,134	,003
,6765323	,104	,003
,7257027	,090	,003
,7626773	,075	,003
,8250212	,060	,000
,8791760	,015	,000
1,0000000	,000	,000

4.5. Discusión Final

Después de haber establecido las justificaciones resultantes de la investigación es de precisar que el modelo ha sido diseñado con todas las estrategias de análisis cuantitativo con respecto a la metodología estadística que le pertenece a la regresión logística binaria y que está fundamentada en la parte teórica de esta técnica, se ha considerado realizar el procedimiento bajo las estrictas medidas de confiabilidad que exige el análisis multivariante, aquello para no obviar detalles de carácter particular en los datos utilizados como respuestas de las personas encuestadas que influyan en el resultado; en el modelo que se trataba de diseñar,

en nuestro caso los factores fueron resumidos en un solo bloque por tema de capacidad de información y redundancia de las respuestas que se consideraban en la base de datos original; ya que solo se consideran aquellos factores que intervienen de manera propia en la investigación, cumpliendo así uno de los supuestos importantes de la regresión logística, al final se concluye resultados muy específicos en cuanto a lo que se esperaba encontrar, puesto que en base del conocimiento empírico desarrollado en temas de salud; como nuestro estudio se sabe con anterioridad de la existencia de factores en el comportamiento de las personas que influyen para que dicha persona desarrolle cualquiera de estas enfermedades, ahora en la parte de resultados finales dado que el número de variables necesarias encontradas en los modelos son pocas en comparación con las iniciales, es correcto afirmar que la evaluación de nuestro estudio debe ser utilizado como un ejemplo base que sirva de guía para investigaciones similares, además este trabajo no tiene forma de establecer alguna comparación semejante con un estudio similar, en nuestro caso presentar esta investigación es el paso inicial para lograr el objetivo real de nuestra especialidad, que es la aplicación y utilización de modelos de clasificación como un proceso de manejo continuo en procedimientos y evaluaciones de salud, teniendo el modelo encontrado como fundamento para investigaciones futuras y más exhaustivas en relación con los principios médicos específicos que se deben considerar en este tipo de temas de salud. Sin dejar de aseverar que el modelo estadístico proporcionado es sin duda una herramienta principal de utilización para este tipo de investigaciones y aunque no exista un trabajo con esas características en la actualidad cercana para comparar, se sabe que puede ser seleccionada como la mejor decisión en el conjunto de posibilidades estadísticas resumidas a priori, por conocimiento del tipo de variables, los elementos y objetivos que intervienen en este tipo de procedimientos.

V.CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

1. El análisis de Regresión Logística Binaria (Modelo Logit Dicotómico Multivariado) es una técnica estadística donde el procedimiento que está incluido en dicha técnica para este tipo de información y en base las variables que intervienen a esta investigación se ajustan al comportamiento de la población del departamento de Piura, de acuerdo a la información que se tiene registrada en las bases de datos de la institución de la cual procede la iniciativa de realizar este estudio, por tal motivo esta técnica sirve para predecir la existencia de las enfermedades que se han considerado, eso según la bondad de ajuste del modelo, aunque se debe ser consciente que los factores implicados y que resultaron significativos son pocos en comparación con lo que se desearía haber encontrado según la literatura en medicina que se consultó de forma previa respecto de dichas enfermedades.

2. Las Variables que resultaron estadísticamente significativas para el cálculo de la Probabilidad de contraer mediante ambos modelos de regresión logística binaria (Logit Dicotómico Multivariado) son: Las variables que incrementan la probabilidad de la existencia de la enfermedad diabetes, según la descripción de la tabla encontrada en los resultados es solo 1: presencia de Hipertensión en la persona. Por otro lado las variables que disminuyen la probabilidad de la existencia de la enfermedad diabetes son 2: Nivel de esfuerzo Físico, Consumo de frutas estas dos últimas variables presentan un efecto negativo según el cuadro mostrado.

Ecuación 1: expresión de la ecuación del modelo para la Diabetes.

$$p = \frac{1}{1 + e^{-z}}$$

Dónde:

$$Z = -0.181 + 2.030 * \text{Pre\_Hp} - 1.204 * \text{Niv\_Esf} - 1.100\ \text{Cons\_Frut}$$

Las variables que incrementan la probabilidad de la existencia de la enfermedad Hipertensión, según la descripción anterior son solo 2: presencia de Diabetes en las personas y además realización de deporte por parte de los encuestados. Por otro lado las variables que disminuyen la probabilidad de la existencia de la enfermedad Hipertensión son también 2: Presencia de Colesterol y número de horas en ver televisión.

Ecuación 2: expresión de la ecuación del modelo para la Hipertensión.

$$p = \frac{1}{1 + e^{-Z}}$$

Dónde:

$$Z = 2.611 + 1.810 * Pres_{Diab} + 0.599 * Real_{Dep} - 2.103\, Pres_{Col} - 0.819\, Num\_HTel$$

Del cual sus coeficientes β_i son estadísticamente significativos al nivel del 5%. Los modelos se ajustan correctamente de acuerdo a la prueba de Bondad de Ajuste de Hosmer y Lemeshow, realizada anteriormente en ambos casos. Ambas ecuaciones son las que se utilizaran para el cálculo de probabilidades en dichas enfermedades, para predecir la existencia o no en las personas encuestadas o en el futuro como modelo de pronóstico.

3. Para interpretar de forma sencilla, se debe recordar que el modelo original esta en términos del logaritmo natural de las oportunidades o logit. Por lo tanto, el coeficiente B es el efecto de una unidad de cambio en una variable independiente sobre el logaritmo natural de las oportunidades. El significado real en términos de la probabilidad, es lo que nos interesa llegar a conocer de forma más intuitiva y esto se deduce en la columna Exp (B) de las tablas de variables incluidas en los modelos. Este se expresa ahora en términos de la razón de oportunidad sobre la variable dependiente, los factores implicados en el primer modelo de la enfermedad Diabetes son los siguientes con su respectiva razón de oportunidad o también llamado teóricamente en bioestadística como Odds ratio y se mencionan a continuación: Presencia de Hipertensión en la persona con 7.610, Nivel de

esfuerzo Físico con 0.300, Consumo de frutas con 0.333 estas dos últimas variables presentan un efecto inferior a la unidad según el cuadro mostrado y la interpretación se realizara de acuerdo a la característica de la codificación de la variable, los factores del segundo modelo establecido para la existencia de la enfermedad Hipertensión, según la descripción anterior, con sus respectivas razón de oportunidad son las siguientes: presencia de Diabetes en las personas con 6.111 además realización de deporte por parte de los encuestados con 1.821, por otro lado Presencia de Colesterol con 0.122 y número de horas en ver televisión con 0.441, teniendo la misma consideración del modelo anterior en cuanto a su interpretación.

4. Las tablas de clasificación muestran que el porcentaje global de predicciones correctas es 91.40% en el modelo, para la primera enfermedad estudiada llamada Diabetes; se observa también que el número de pacientes que presentan la enfermedad son pronosticados correctamente en 52.9% (sensibilidad) y del total de personas que no tiene la enfermedad de estudio el 93.1% (especificidad) se pronostican correctamente. La tabla de clasificación siguiente muestra que el porcentaje global de predicciones correctas es 84.3% en el modelo, para la segunda enfermedad estudiada llamada Hipertensión; se observa también que el número de pacientes que presentan la enfermedad son pronosticados correctamente en el 44.8% (sensibilidad) y del total de personas que no tiene la enfermedad de estudio el 91.8% (especificidad) se pronostican correctamente, logrando así establecer valores que respondan al objetivo planteado con respecto a la capacidad predictiva del ambos modelos.

5. Mediante la Técnica estadística de la Curva COR se ha determinado que en el primer modelo de la enfermedad llamada Diabetes, el punto óptimo de corte sería 0.068235 es el resultado de capacidad predictiva que proporciona un resultado distinto algo mejor en función de la categoría de interés, dado que se obtiene una Sensibilidad de 58.8% pero una reducción en la Especificidad con un valor de 87.3%. Además en el cuadro siguiente se puede observar que para el segundo modelo de la enfermedad llamada Hipertensión, el punto óptimo de corte es 0.217475 es el resultado de capacidad predictiva que proporciona un resultado distinto algo mejor en función de la categoría de interés, dado que se obtiene una

Sensibilidad de 52.2% pero una reducción en la Especificidad con un valor de 85.6%. Esto hace suponer que se necesita un modelo de análisis igual en el futuro con una muestra distinta y tener así una comparación similar para establecer conclusiones, en función de los puntos de corte elegidos.

6. El modelo que se ha determinado en esta investigación, si bien es cierto cumple con los indicadores necesarios para concluir que tiene una buena capacidad predictiva desde el punto de vista estadístico, es necesario aun efectuar algunos ajustes e incluir algunas variables con las que cuenta la institución en sus bases de datos, a las cuales no se ha tenido acceso para efectos de esta investigación. Por tal motivo la determinación de este modelo constituye un gran paso inicial para alcanzar el objetivo de la institución de utilizar modelos de pronósticos más avanzados para calcular probabilidades acerca de la existencia de estas enfermedades en Poblaciones específicas como nuestro departamento, aunque debe ser sometido a juicio de expertos y otras metodologías análisis que ayuden a mejorar los resultados que hemos obtenido en esta investigación.

5.2. RECOMENDACIONES

Con base en lo anterior, en seguida se enumera una serie de recomendaciones:

1. En nuestro trabajo es preciso mencionar que no proponemos factores que disminuyan estas enfermedades, puesto que por conocimiento médico y de salud ya están establecidos para lograr dicha reducción, este trabajo solo pretende diseñar un modelo que dé a conocer cuáles son los factores estadísticamente comprobables que influyan en estas enfermedades, y por eso se recomienda no concluir una expresión medica porque no es el objetivo principal de esta investigación.

2. Utilizar herramientas estadísticas de tipo exploratorio para analizar minuciosamente la información con la que se cuenta.

3. Detectar la relación entre las enfermedades y los distintos factores que influyen de manera negativa en cada uno de estos problemas estudiados, a través de la razón de oportunidad de los factores significativos encontrados con el modelo logístico.

4. La investigación aplicada a temas de la Salud, implica primero un estudio unidimensional de forma anticipada o previa en cada variable o factor que se debe incluir en el análisis de regresión logística binaria, porque proporciona una conclusión inicial de la relación que existe entre ellas y la variable de Respuesta. A través de la prueba estadística Chi-Cuadrado que nos da la significancia acerca de la influencia de dichos factores.

5. Las bases de datos de la institución están clasificadas en preguntas acerca del comportamiento de las personas frente a los factores de salud asociados a estas enfermedades, no existe una categorización estandarizada en las posibles respuestas de los encuestados, por tal motivo se recomienda analizar cuidadosamente cada factor para poder interpretar de forma eficiente los indicadores estadísticos.

6. Se puede rehacer la investigación con una reclasificación de los factores en nuevas variables en categorías uniformes para todos ellos, de acuerdo a la variable dependiente y lograr una comprensión más sencilla de los resultados, esta recomendación también es extensible para proyectos similares en el futuro bajo las mismas características, con información original como la que nos proporcionó el INEI.

VI. BIBLIOGRAFÍA

1. Buchma A, Boyle P, Wilson R, Fleischman D, Leurgans S, Bennett D, **Association Between Late-Life Social Activity and Motor Decline in Older Adults.** Archives of Internal Medicine [Internet]. Jun 2009 [citado 17 Enero 2010]; 169 (12): 1139-1146. Disponible en la World WideWeb:http://archinte.amaassn.org/cgi/content/short/169/12/1139#otherarticles

2. Hernández, R.; Fernández, C. y Baptista, P. (2008). Metodología de la Investigación (4ª Edición). México: McGraw-Hill.

3. Johnson, D. (2000). Métodos Multivariados Aplicados al Análisis de Datos (2ª Edición). México: Thompson Editores Internacional.

4. Goicochea Rios E, Chian Garcia A. Estado de Salud de los Adultos Mayores en el Servicio 15 de la Cartera de Atención Primaria del Hospital I Albrecht – EsSalud. Noviembre 2006 a Julio 2007. Revista Salud, Sexualidad y Sociedad. 2008 [citado 18 de Enero 2010]; 2 (1). Disponible en la World Wide Web: http://www.inppares.org.pe/revistasss/Revista%20II%202009/7%20-%20Adultos%20Mayores.pdf

5. Manrique Betty, Salinas Aarón, Téllez Martha. **Factores asociados con la dependencia funcional en los adultos mayores beneficiarios del Programa Oportunidades.** 2008 [citado 18 de Enero 2010]. Disponible en la World Wide Web: http://www.alapop.org/2009/images/DOCSFINAIS_PDF/ALAP_2008_FINAL_278.pdf

6. Llibre Guerra J, Guerra Hernández M, Perera Miniet E. **Comportamiento de las enfermedades crónicas no trasmisibles en adultos mayores** [Internet]. 2008 [citado 11 de Enero 2010]; Disponible en la World Wide Web: http://www.bvs.sld.cu/revistas/mgi/vol24_4_08/mgi05408.htm

7. Medina, E. (2003). Modelos De Elección Discreta. Investigación Desarrollada en la Universidad de España. Disponible en: http://www.uam.es/personal_pdi/economicas/eva/pdf/logit.pdf

8. Manual de la entrevistadora – ENDES, 2013.

9. Ordaz, J.; Melgar, M. y Rubio, C. (2008). Métodos Estadísticos y Econométricos en la Empresa Para Finanzas. Universidad Pablo de Olavide. Disponible en: http://www.upo.es/export/portal/com/bin/portal/upo/profesores/jaordsan/profesor/1328642345406_metodos_estadisticos_y_econometricos_en_la_empresa_y_para_finanzas.pdf

10. Organización Mundial de la Salud: Informe sobre la Situación Mundial de las Enfermedades No Transmisibles, 2010.

11. Perú-Ministerio de salud/Dirección General de Epidemiología. "Análisis de la Situación de Salud del Perú", 2010.

12. Quiroga, G. y García, R. (2006). Calculo del Riesgo del Microcrédito en Operaciones de Finagro para una Entidad Financiera del Eje Cafetero. Disponible en: http://www.bdigital.unal.edu.co/983/1/germanalonsoquirogazapata.2006.pdf

13. Rayo, S.; Lara, R. y Camino, D. (2010). Un Modelo Credit Scoring para Instituciones de Microfinanzas en el Marco de Basilea II. Disponible en: http://www.esan.edu.pe/publicaciones/2010/06/02/05.pdf

14. Romero Cabrera **A.** Perspectivas **actuales en la asistencia sanitaria al adulto mayor.** Rev Panam Salud Pública. 2008; 24(4): 288 – 94.

15. Ruíz Dioses L, Campos León M, Peña N. Situación sociofamiliar, valoración funcional y enfermedades prevalentes del adulto mayor que acude a establecimientos del primer nivel de atención, Callao 2006. Rev. Perú. Med. Exp. Salud Pública [Internet]. Oct./Dic. 2008 [citado 11 de Enero 2010]; 25 (4), p.374-379. Disponible en la World WideWeb:

http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1726-46342008000400005&lng=es&nrm=iso

16. Uriel, E. y Aldas, J. (2005). Análisis Multivariante Aplicado – Aplicaciones al Marketing, Investigación de Mercados, Economía, Dirección de Empresas y Turismo (1ª Edición). México: Thompson Editores Internacional.

17. Villariño, A. (2010). La Gestión del Riesgo de Crédito. Socio Consultor de MDV CONSULTORES – España. Disponible en: http://www.angelvila.eu/Publicaciones_PDF/Gestion_Riesgo_Credito.pdf

VII. ANEXOS DE LA INVESTIGACIÓN

Anexo 01

Los cuadros que se deben presentar en esta parte corresponden a aquellos valores indicados en el capítulo de resultados, que han sido utilizados para encontrar la diferencia en el indicador de verosimilitud de ambos modelos.

Cuadro 7.1: Variables en la ecuación para el Modelo de la Diabetes en el bloque inicial.

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-3,166	,248	163,475	1	,000	,042

Cuadro 7.2: Historial de Iteraciones del Modelo de la Diabetes en el bloque inicial.

Iteración		-2 log de la verosimilitud	Coeficientes Constante
Paso 0	1	186,531	-1,838
	2	147,328	-2,655
	3	142,509	-3,066
	4	142,342	-3,161
	5	142,342	-3,166
	6	142,342	-3,166

Cuadro 7.3: Resumen de las variables que no están en la ecuación en el bloqueo inicial.

			Puntuación	gl	Sig.
Paso 0	Variables	Presencia Colesterol	11,510	1	,001
		Presencia_hipertension_Alta	24,287	1	,000
		Tipo Actividad	4,426	1	,035
		Nivel_Esfuerzo_Fisico	9,178	1	,002
		Realizacion_Deporte	,968	1	,325
		Numero_horas_television	,037	1	,846
		Consume_Sal_Comida	,543	1	,461
		Consume Ensaladas	,758	1	,384
		Consume_Frutas	4,908	1	,027
		Consume Dulces	1,611	1	,204
		Numero_Dias_Fritura	2,102	1	,147
		Alimentacion_sin_Grasas	6,091	1	,014
		Consumo Cigarrillos	,557	1	,455
		Consume_Bebidas_Alcoholicas	,886	1	,346
	Estadísticos globales		43,253	14	,000

En este siguiente cuadro se da a conocer los factores que no interviene en el modelo inicial, este cuadro muestra cuales son las variables que sin análisis estadístico, pueden propiciar significancia en la predicción.

Cuadro 7.4: Resumen de las variables que no están en la ecuación en el bloqueo final.

			Puntuación	gl	Sig.
Paso 3	Variables	Presencia Colesterol	1,261	1	,261
		Tipo Actividad	1,408	1	,235
		Realizacion_Deporte	,008	1	,930
		Numero_horas_television	,071	1	,790
		Consume_Sal_Comida	,024	1	,877
		consume Ensaladas	,103	1	,748
		Consume_Dulces	3,173	1	,075
		Numero_Dias_Fritura	,396	1	,529
		Alimentacion_sin_Grasas	3,159	1	,075
		Consumo_Cigarrillos	,039	1	,843
		Consume_Bebidas_Alcoholicas	,349	1	,554
	Estadísticos globales		8,879	11	,633

En este siguiente cuadro se da a conocer los factores que no interviene en el modelo final, este cuadro muestra cuales son las variables que con el análisis estadístico, ya realizado no pueden propiciar significancia en la predicción.

Cuadro 7.5: Cuadro resumen de los pasos para la Diabetes.

	Mejora			Modelo			% de clas. correcta	Variable
Paso	Chi cuadrado	gl	Sig.	Chi cuadrado	gl	Sig.		
1	17,124	1	,000	17,124	1	,000	84,8%	IN: Presencia_hipertension_Alta
2	5,425	1	,020	22,549	2	,000	93,1%	IN: Nivel_Esfuerzo_Fisico
3	4,313	1	,038	26,863	3	,000	91,4%	IN: Consume_Frutas

Cuadro 7.6: Variables en la ecuación para el Modelo de la Hipertensión en el bloque inicial.

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-1,662	,133	155,505	1	,000	,190

Cuadro 7.7: Historial de Iteraciones del Modelo de la Hipertensión en el bloque inicial.

Iteración		-2 log de la verosimilitud	Coeficientes Constante
Paso 0	1	374,073	-1,362
	2	368,698	-1,635
	3	368,659	-1,662
	4	368,659	-1,662

Cuadro 7.8: Resumen de las variables que no están en la ecuación en el bloqueo inicial.

		Puntuación	gl	Sig.
Paso 0 Variables	Presencia_Diabetes	24,287	1	,000
	Presencia_Colesterol	48,548	1	,000
	Tipo_Actividad	1,879	1	,170
	Nivel_Esfuerzo_Fisico	5,741	1	,017
	Realizacion_Deporte	6,252	1	,012
	Numero_horas_television	3,207	1	,073
	Consume_Sal_Comida	,701	1	,403
	Consume_Ensaladas	,471	1	,492
	Consume_Frutas	,025	1	,874
	Consume_Dulces	,079	1	,779
	Numero_Dias_Fritura	5,411	1	,020
	Alimentacion_sin_Grasas	10,518	1	,001
	Consumo_Cigarrillos	,161	1	,688
	Consume_Bebidas_Alcoholicas	,062	1	,803
Estadísticos globales		79,986	14	,000

En este siguiente cuadro se da a conocer los factores que no interviene en el modelo inicial, este cuadro muestra cuales son las variables que sin análisis estadístico, pueden propiciar significancia en la predicción.

Cuadro 7.9: Resumen de las variables que no están en la ecuación en el bloqueo final.

		Puntuación	gl	Sig.
Paso 4 Variables	Tipo_Actividad	,160	1	,689
	Nivel_Esfuerzo_Fisico	1,141	1	,285
	Consume_Sal_Comida	,001	1	,975
	Consume_Ensaladas	,288	1	,591
	Consume_Frutas	,210	1	,647
	Consume_Dulces	,039	1	,844
	Numero_Dias_Fritura	,904	1	,342
	Alimentacion_sin_Grasas	3,549	1	,060
	Consumo_Cigarrillos	,077	1	,781
	Consume_Bebidas_Alcoholicas	,847	1	,357
Estadísticos globales		6,717	10	,752

En este siguiente cuadro se da a conocer los factores que no interviene en el modelo final, este cuadro muestra cuales son las variables que con el análisis estadístico, ya realizado no pueden propiciar significancia en la predicción.

Cuadro 7.10: Cuadro resumen de los pasos para la Hipertensión.

	Mejora			Modelo				
Paso	Chi cuadrado	gl	Sig.	Chi cuadrado	gl	Sig.	% de clas. correcta	Variable
1	37,935	1	,000	37,935	1	,000	83,3%	IN: Presencia Colesterol
2	10,850	1	,001	48,785	2	,000	82,9%	IN: Presencia Diabetes
3	9,076	1	,003	57,861	3	,000	84,5%	IN: Numero_horas_television
4	4,092	1	,043	61,953	4	,000	84,3%	IN: Realizacion_Deporte